DOCUMENT RESUME

ED 084 292 TM 003 300

AUTHOR Proger, Barton B.; Mann, Lester
TITLE An Historical Review of Theoretical and Experimental Literature on the Teaching Values of Informal (Nonstandardized), Teacher-Made Achievement Tests: 1913 - 1968.
INSTITUTION Montgomery County Intermediate Unit 23, Blue Bell, Pa.
PUB DATE Oct 73
NOTE 135p.

EDRS PRICE MF-$0.65 HC-$6.58
DESCRIPTORS *Achievement Tests; Feedback; Grades (Scholastic); *Instructional Aids; *Literature Reviews; Student Attitudes; *Teacher Developed Materials; Testing; Values

ABSTRACT

A study of informal, teacher-generated testing activities is presented. The following major topics are covered: (1) frequency of informal achievement testing as related to test learning (2) informal achievement test grades in relation to testing as a learning device, (3) test correction with respect to informal achievement testing as a learning device, (4) test result feedback as related to testing as a learning device, (5) pretesting as an aspect of testing as a learning device, (6) retesting as an aspect of testing as a learning device, (7) test expectation as an aspect of testing as a learning device, (8) test exemption as an aspect of testing as a learning device, (9) student preparation for tests as an aspect of testing as a learning device, (10) student attitudes toward informal achievement tests, (11) test type as an aspect of testing as a learning device, and (12) "test-like events" as an aspect of testing as a learning device. (CK)

An Historical Review Of Theoretical and Experimental Literature On The Teaching Values Of Informal (Nonstandardized), Teacher-Made Achievement Tests: 1913 - 1968

by

Barton B. Proger and Lester Mann

Montgomery County Intermediate Unit No. 23

October 1973

Montgomery County Intermediate Unit No. 23
Special Education Center
1605-B West Main Street
Norristown, Pennsylvania 19401

Preface

Recently several reviews of testing practices have appeared (e.g., Marjorie C. Kirkland, "The Effects of Tests on Students and Schools," Review of Educational Research, 1971, 41, 303-350; Richard C. Anderson, "How to Construct Achievement Tests to Assess Comprehension," Review of Educational Research, 1972, 42, 145-170; Barton B. Proger and Lester Mann, "Criterion - Referenced Measurement: The World of Gray versus Black and White," Journal of Learning Disabilities, 1973, 6, 72-84.) The reviews have emphasized either standardized tests or criterion-referenced measurement. Such topics are receiving the greatest amount of attention from test g experts at present. However, before the advent of tests used in either a norm-referenced measurement (NRM) or a criterion-referenced measurement (CRM) manner, teachers were forced to construct their own, informal devices to assess progress. The reviewers feel that informal, teacher-made tests do not legitimately fall into either the NRM or CRM categories but rather form a third category of their own. It is unfortunate that reviewers of educational research have largely neglected the vast literature on informal, teacher-made tests. At the very least, these studies are of interest from an historical perspective, in that the seeds for many of the ideas behind NRM and CRM were first sown on the informal teacher test domain. This review covers the time period from 1913 to 1968 and thus includes the bulk of exposition on informal, teacher-made tests, since interest in the NRM and CRM movements superceded the former type of tests in the late 1940's.

This review is limited to only those articles of either experimental nature or of philosophical/theoretical nature that relate to the instructional benefits of tests. (The introductory chapter fully explains the premises behind the reviewers' perspective on the teaching values of informal, teacher-made tests.) Needless to say, a great deal has been written about the need to check student progress by means of teacher-made tests, but most of this literature is based only on personal biases of the writers and not on evidence. Thus, the reviewers have set minimal criteria that the studies to be included must include empirical evidence to support the assertions that teacher-made tests are beneficial to the children, or, in lieu of such evidence, must at least contain rational psychological learning theory. Twelve topics eventually delineated themselves: (1) frequency of testing (34 references); (2) test grading (7 references); (3) test correction modes (11 references); (4) test result feedback (22 references);

(5) pretesting (36 references); (6) retesting (9 references); (7) test expectation (7 references); (8) test exemption (13 references); (9) student preparation modes (16 references); (10) student attitudes toward tests (6 references); (11) test type (5 references); and (12) "test-like events" (19 references). In total 185 references were summarized.

This project was completed in connection with several testing research studies carried out by the reviewers and their colleagues from 1967 to the present. The reviewers hope this document will prove useful to others in understanding a somewhat different component to measurement heritage than is commonly recognized in CRM and NRM.

ACKNOWLEDGEMENTS

The reviewers completed this document under the joint auspices of Montgomery County Intermediate Unit No. 23 and one of the federal projects it sponsors: Research and Information Services for Education (RISE; Grant No. OEG-1-67-3010-2696).

The work presented herein was performed pursuant to a Title III Grant (Elementary and Secondary Education Act of 1965) from the Pennsylvania Department of Education acting as the State Educational Agency for the United States Office of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of PDE or USOE, and no official endorsement by either should be inferred. The views expressed are solely those of the authors themselves.

This review was completed in July, 1968, while the senior author was a United States Office of Education Pre-doctoral Fellow in the Educational Research Training Program at Lehigh University (OEG-1-6-06-061757 - 0939, Project 6-1757). Thanks are expressed to Dr. Merle W. Tate, Professor Emeritus of Education at Lehigh, for reading a draft of this review. Thanks are also extended to Charles F. Haughey, former Director of RISE, for making possible the arrangements to complete this review. Finally, the reviewers express deep appreciation to Mrs. Claude H. Bressler, the senior reviewer's sister, for typing the manuscript.

TABLE OF CONTENTS

INTRODUCTION

In no case will the review concern itself with standardized testing; rather, only informal, teacher-generated testing activities are dealt with. The reason for this procedure is justifiable. Several studies have dealt with standardized tests, especially in the contexts of pretesting, retesting, coaching, and so on, with the same instrument, but perhaps with different parallel forms. In other words, these investigators of standardized achievement tests attempted to see what learning takes place with standardized tests themselves (as compared to the learning that takes place in the usual nontesting, lecture aspect of instruction); some would label such standardized test learning benefits under the somewhat undesirable-sounding names of "practice effects", "coaching effects", and so on. Such psychological effects involved in standardized testing are important. However, in the real school situation, standardized achievement tests are given rather infrequently during the school year to any one student. Hence, the practical use to which such experimental conclusions could be put is quite limited indeed.

On the other hand, the informal achievement tests given by teachers throughout the course of instruction constitute a major part of the curriculum. If learning effects (not just the usual evaluative functions) above and beyond the in-class, nontesting, instructional process can be produced from the taking of informal achievement tests themselves, or from using informal achievement tests in a specific way, then such knowledge would be highly practical for the realistic

classroom setting. Hence, this review will direct itself to identifying the various ways in which informal achievement tests can aid instruction. The review is therefore making a unique contribution to testing literature; most testing reviews have considered only the usually cited function of tests: to evaluate and rank the student's achievement. Other testing reviews have dealt with the technical issues of test construction: reliability, validity, item difficulty, item discrimination power, and so on. On the other hand, this review neglects the already well-documented, usually-discussed topics of testing and concentrates only on how a student can actually learn from the very taking itself of tests.

There are several ways in which informal achievement tests can be used to yield learning benefits above and beyond the usual instructional, nontesting part of the classroom procedure. However, running throughout all of the various methods of informal test use is the common thread of the emotion-producing situation of being under the threat of a test. As the reader will see, apparently the threat of a test is psychologically effective enough (albeit in many different ways) to force the test taker to be more careful of the way in which he processes the test information as compared to being in a mere nontesting, practice situation. (Analogies can be made with the results of the voluminous research in programed instruction, although such research has been treated adequately elsewhere. Hence, this review will not concern itself with programed instruction.) In effect, the student under the testing condition perhaps is being forced to concentrate on the material presented in the test moreso than he would under just

a mere practice condition. The clearest example is the essay test as used in high schools and colleges. Here the student is asked to synthesize information he has learned in class in ways somewhat different from those in which the material was originally presented in class. It is reasoned that if the content of the test items is presented in a slightly different order or context from the original presentation during the nontesting, instructional part of the class period, then the student is forced to think about the subject matter in a more meaningful fashion than by mere rote recall. In effect, the content is structured more meaningfully in the student's mind.

The same features of enforced activity with the subject matter of the test questions and the potential structuring effects in the student's mind of such subject matter can be found in all types of tests in all subject areas, not just the essay test that is used in predominantly verbal subjects as compared to science and mathematics. To identify just what it is about a test (the emotional effects, the structuring effects, the practice effects, and so on) that causes increased learning is still a moot point; adequate measuring procedures have not yet been devised, especially for the physiological aspects of test behavior. The reviewer is aware of no studies that have investigated what happens to blood pressure, pulse, brain wave patterns, and so on, in realistic testing situations. It is true that such physiological measures have been taken in unrealistic laboratory situations with respect to rather contrived and often trivial tasks, but these studies do not concern the topic at hand.

Nonetheless, certain aspects of informal, realistic achievement tests as a learning device have been identified and manipulated in an effort to yield learning benefits above and beyond the usual non-testing, instructional part of the procedure: (1) frequency of testing, (2) test grades, (3) test correction, (4) test result feedback, (5) pre-testing, (6) retesting, (7) test expectation, (8) test exemption, (9) student preparation for tests, (10) student attitudes toward tests, (11) test type, and (12) "test-like events". Before reviewing the corresponding references for each of the twelve topics, a brief description of each area will be given in the context with which it will be used throughout the review.

The first topic, frequency of testing, can be found to aid learning. In this review, frequency of testing is defined as how often the teacher gives an informal achievement test in the course of instruction. When one controls all other pertinent variables, he can easily see how frequency of testing might at least create the potential for increased learning beyond the usual in-class instructional procedure. First, the students, by the very enforced activity of going through the material on the test, are getting additional practice with the subject matter. Second, the threat of frequent tests might motivate the students to prepare their lessons better outside of class. Third, the students probably gain insight from the tests as to which topics in the subject matter are most important and therefore should be mastered.

Test grades, the second topic in this review, will involve only

the use of so-called extrinsic motivation in the form of grades received on informal achievement tests. (Although much research and exposition already exists on systems of report card grading, such motivation is outside the periphery of this paper.) If students know what grade they received on an informal achievement test, this situation might be expected to motivate them to greater accomplishment in later performance. Any rivalry that might build up between students in a class could also have a beneficial motivating effect.

The third topic of this review, test correction, concerns whether or not the students correct their informal achievement tests in class, as well as whether or not the teacher provides comments about the mistakes. It might be expected that, above and beyond the learning that takes place during the nontesting instructional process itself, a student can learn additional information from the way in which mistakes are corrected. First, if the tests are randomly handed back to the students the next day while the test format and instructional process content are still clear in the students' minds, it would logically be expected that additional learning will take place when the student sees the errors of others, asks himself why a problem is wrong, and tries to relate this to what he did on his test. Second, regardless of this first consideration, after correction of the tests (either by the teacher or by the students), if the teacher then writes comments on each paper pointing out a student's difficulties, or praising him, additional learning would again logically be expected to occur. Under different conditions, such additional learning might not be expected to occur.

Test result feedback, the fourth topic of this review, concerns whether or not the teacher gives the corrected tests back to the students as well as whether or not the teacher discusses the errors in class. This is perhaps one of the most fertile areas from which test learning benefits can be derived. If the student receives his corrected informal achievement test back as soon as possible, it might logically be expected that he will still be interested enough to examine whatever errors he had on the test. The student might analyze just why he made such errors and how he might rectify them. Further, if the teacher also discusses the general classes of errors made in the test, the students would logically be expected to benefit from such a discussion.

The fifth topic of this review, pretesting, deals with the effect of giving a pretest over the unit of instruction to be studied before such study is actually begun. This is a particularly interesting topic, since it involves not only psychological problems but also methodological ones as well. Psychologically, if students are given a pretest over the unit of instruction to be studied, then one might expect the subject matter of the future instructional unit to be structured to some extent in their minds; the students know what to look for in their ensuing study by the very nature of the questions asked on the pretest itself. Methodologically, several investigators have considered such learning benefits in a negative light and called them "test sensitization"; these investigators are interested mainly in the "practice effects" that were mentioned earlier.

Retesting, the sixth topic of this review, will concern itself

with the phenomena of recall and reminiscence in connection with realistic, meaningful classroom material. If students are retested (as compared to mere once-and-done posttesting) on subject matter, they might be expected to retain it longer than those who are not retested at periodic intervals. It should be noted that retesting deals with using the same instrument (or an equivalent form) over and over with respect to the same specific subject material, while the earlier-mentioned frequency of testing concerns itself with using different instruments covering different specific subject matter in a high-frequency schedule of testing. Practice effects and structuring effects are probably pertinent issues in retesting.

The seventh topic of this review, test expectation, deals with whether or not the students have been warned of an approaching test. If one is warned about a future test, he might be expected to study more than he ordinarily would outside of class in preparation for the test. On the other hand, certain students who are affected adversely by the very concept of "test" might do better if they are given the test in an unannounced fashion, relying on their usual, noncramming study habits; this might logically be expected to be true in the case of poor achievers.

Test exemption, the eighth topic of this review, concerns itself with both exemption from testing and exemption by testing. Both have motivational properties and can be expected to increase learning above and beyond the usual level of direct, in-class learning. If a student knows he can avoid certain tests by demonstrating a certain level of

competence in his daily practice work, or if he knows he can avoid repetitious work by taking an examination on it, then he might be expected to exert greater effort during the usual, in-class instructional process.

The ninth topic of this review, student preparation for tests, deals mainly with the type of test the student expects. If he anticipates a test that emphasizes very specific details, he might be expected to study in a different manner than if he expects a test dealing with broad generalities.

Student attitudes toward informal achievement tests, the tenth topic of the review, concerns any systematic survey into students' preferences for various informal testing procedures. This is self-explanatory.

The eleventh topic of this review, test type, deals with attempts to determine whether or not different test types (multiple-choice, completion, true-false, essay, and so on) used in connection with the same specific subject matter will yield differential learning benefits as measured by follow-up uniform testing procedures. This can be a very significant topic for the actual classroom situation. For example, if a student can be forced to synthesize and process subject matter more effectively on one type of test as compared to other types, then the teacher would do well to use such a test type frequently. However, it must be noted here that the reviewer is interested only in learning benefits that accrue to the student; the advantages and disadvantages of one test type versus another

with respect to technical test features (reliability, validity, difficulty, and so on), administrative efficiency, and so on have been adequately covered elsewhere by specialists in tests and measurements.

The twelfth and final topic of this review, "test-like events", actually forms a convenient bridge to the discussion on implications for further research. In fact, the topic of "test-like events" has formed the intensive and recent research efforts of only a few experts in the field of learning theory. The phrase "test-like events" was coined by Ernst Z. Rothkopf of Bell Telephone Laboratories, Inc., to cover learning situations using written, highly verbal, and non-programed material (that is, the most commonly used expository passages used in such courses as English, history, geography, and so on) where the student is evaluated frequently by means of study questions in a "test-like" (that is, evaluative) manner but yet not a true testing situation. Further, Rothkopf has coined the term "mathemagenic behavior" to cover all the emotional, physiological, and cognitive activities the student engages in as he learns the written material via the study questions. The techniques of investigation used by Rothkopf form a unique methodology for investigating the other eleven topics mentioned above that was never available before. Thus, the fine points of the testing situation (structuring effects, practice effects, and so on) that produce additional learning benefits can at last be investigated to a depth never achieved before. One might conclude that any type of questioning activity could be labeled "test-like"; however, this review will consider the effects of study questions only as they relate to

written material such as textbook passages (and the extension of such written passages into the real-life, written, test situations). Other questioning activities, such as oral questioning and recitation in class, homework for homework's sake outside of class, and so on, will not be considered as truly "test-like" in nature. Only those nontesting, written, questioning activities in reference to corresponding written passages are considered to be sufficiently "test-like" in nature to warrant inclusion in this review of test learning benefits.

The preceding discussion completes a brief overview and informal definition of each of the twelve topics to be taken up in this review. However, before attempting a review of the first topic, frequency of testing, it will aid the reader if he first considers the general, nonresearch references that have suggested the additional learning benefits above and beyond the usual nontesting aspects of the instructional process that can arise from using informal achievement tests in specific ways.

INTRODUCTORY REVIEW OF THEORETICAL REFERENCES ON INFORMAL ACHIEVEMENT TEST LEARNING BENEFITS

McKeachie (1963, p. 1154) says:

> While we usually think of testing procedures in terms of their validity as measures of student achievement, their function as instruments for promoting learning may be even more important. After dismal recitals of nonsignificant differences between different teaching methods, it is refreshing to find positive results from variations in testing procedures.

Anderson (1960, p. 50) provides a well-stated summary of the whole problem:

> But to say that teachers don't sometimes begrudge the time taken for [informal achievement] testing and that most students face 'test day' with real enthusiasm is going too far in the other direction. Yet this is just what we may be able to say soon if tests can be utilized to support the process students and teachers are most concerned about--if tests can be used to teach students something. Furthermore, it may be the case that eventually tests will be as useful for teaching as for measuring. [underlining inserted by reviewer]

In support of the experiment presently being conducted by the reviewer, Gardner (1953, p. 87) says, ". . . more research should be done regarding practical problems encountered by teachers in the classroom (and by students as well) in their use of both standardized and informal tests." [underlining inserted by reviewer]

Koester (1957) claims that the evaluative function of informal achievement tests is overemphasized in relation to their instructional potential. Making planned use of informal achievement tests in high school English classes, asserts Kimmel (1923), has yielded greater than usual learning benefits, although no actual experiment was carried out. Obourn (1932) cites the opinions of several physics teachers in high school who have found that informal achievement tests can be used as teaching devices as well as evaluative instruments.

Ruch (1929, p. 145) urges that "we must abandon the thoroughly untenable position that time spent in testing is time wasted in teaching. Teaching and testing are aspects of the same process."

Many other early writers and theorists have written about the learning benefits that can arise from informal achievement tests: Butler (1922), Elston (1923), Lockhart (1928), Woody (1929), Fenton (1929), Henricksen (1930), and Symonds (1933). However, none of the references in this section of the review have supplied experimental evidence to support their beliefs.

This completes the introductory review of nonresearch references which have referred to the instructional values of informal achievement tests only in a general way. Nonresearch references that try to hypothesize precisely how such additional learning benefits arise will be treated in the appropriate section in the twelve specific reviews that follow. The first topic is frequency of testing, which is of major importance to the reviewer in his dissertation.

REVIEW TOPIC ONE: FREQUENCY OF INFORMAL ACHIEVEMENT TESTING AS RELATED TO TEST LEARNING BENEFITS

Nonexperimental References: Wrightstone (1963, pp. 50-51) gives those interested in frequency of informal achievement testing some precise guidelines but fails to account for the many contradictory findings:

> Some persons have assumed that more frequent tests will increase the motivation and effort of the student to achieve immediate educational goals. Carried to a ridiculous conclusion, this might mean one test per teaching period. When tests are administered too frequently, their motivational value is reduced. In a variety of fields at the college level, studies show that when weekly tests are given, discussed, and corrected, the lower-ability students achieve more on a final examination of similar questions than with less frequent examinations. The more-able

> students may be retarded because of too frequent testing. The less able profit mainly from direction of their learning to specifics and to practice in selecting the correct responses. The more-able are not aided by frequent--weekly or daily--tests. [underlining inserted by reviewer]

Unfortunately, Wrightstone omitted all negative results in the list of experiments he cites to support his conclusions; further, the "positive" results of the experiments he does cite are often confounded by other uncontrolled variables. Moreover, one still has no concrete evidence as to what occurs in the elementary school where the ideas toward testing in general are being molded in the students' minds; the investigators and theorists continuously emphasize college studies but only rarely touch on the more crucial, formative years: kindergarten through grade twelve.

Gardner (1953, p. 87) displays a common misconception among educational theorists with respect to frequent testing. In relation to one of the experimental studies in this field, he says that the investigators ". . . have again demonstrated the motivating effect of frequent testing." In the first place, not many investigators have "demonstrated" such an effect. To the contrary, many conflicting reports are available; positive results have been the exception, not the rule. It appears that any results that are obtained from frequent informal achievement testing must be qualified with respect to control variables such as grade level, previous achievement, sex, and so on. The practical implications of this misconception are important; no doubt many teachers are presently

laboring under the belief that frequent testing will drive their students on to higher achievement. However, no safe conclusions can be drawn on this issue.

Many early theorists voiced their support of frequent informal achievement tests. Ruch (1929) says infrequent tests of an extensive nature should be laid aside in favor of shorter, more frequent testing. Pearson (1929) supports frequent informal achievement testing in his city school system. Odell (1928) thinks pupil achievement will increase under frequent informal testing but gives no evidence. Further, he suggests that both slow and rapid learners will benefit from a program of frequent testing. Finally, he says that such informal achievement tests should be given often only if they are relatively short. Opdyke (1927) also voiced the latter idea. Parker (1920) says frequent testing is needed to make students prepare their lessons adequately. Ragusa (1930) offers the same opinion and in addition claims that stort objective tests should be given two or three times a week.

Experimental Findings: In this section all of the experiments that the reviewer was able to find in his search of the literature are presented in connection with frequency of testing. The review of experiments will be conducted in three parts: (1) elementary school, (2) junior and senior high school, and (3) post-high school. The experiments will be taken in chronological order.

Each experiment will be described in as much detail as is possible and practical. Particular attention will be given to flaws

in design and analysis. However, before getting into specific criticisms of any one experiment, a general distinction will be made by the reviewer between two types of allegedly "loose" research. The first type will be termed "broad curricular research" by this reviewer for expediency's sake later in this paper. "Broad curricular research" is defined here as an experimental comparison between two curricular programs of instruction. For example, a school district might be interested in comparing the effectiveness of a new laboratory-discovery approach of teaching junior high school science against the traditional lecture-textbook method. No matter how much effort is taken to control extraneous factors, the most that one will be able to conclude from his results is that the new program taken as a whole is or is not more effective from the cumulative achievement standpoint than the traditional program; one will never be able to isolate just what aspect of the new program was or was not a causative factor in the results. In effect, control in the "basic research" sense is lacking. Many potential causative factors are confounded with each other. The above definition, then, is what the reviewer will henceforth mean by "broad curricular research." However, it must not be inferred by the reader that the reviewer looks down upon the above type of research; indeed, such research is a necessary ingredient of curricular progress.

The second type of "loose" research is the "confounded, non-curricular experiment." In the above definition of "broad curricular research" the confounding of various component factors is

unavoidable; these factors are inherent to the curricular program and hence cannot be "controlled" (or isolated) without destroying the integrity of the method. However, the reviewer considers a topic such as frequency of testing to be independently manipulatable of the particular curricular setup being used. In other words, not considering administrative difficulties, a topic such as frequency of testing should be highly amenable to control in the research sense. A large number of experiments, however, have inadvertently confounded the factor of frequency of testing by the manipulation of other variables at the same time as frequency of testing. Thus, in this review, the "confounded, noncurricular experiment" will be considered as a definite design error that could have been avoided by thoughtful planning. (The reader should also be aware, however, that a confounded design can be a deliberately planned advantage--rather than an inadvertent error--if the investigator is interested in highlighting certain interactions or main effects. However, all confounded noncurricular experiments in the following sections of this review were design blunders and not sophisticated analytical refinements).

The reviewer is now ready to proceed with the discussion of experimental studies under the topic of frequency of testing. The first set of experiments to be considered is that of the elementary school.

(a) Elementary School: As already stated previously, the reviewer considers testing procedures in the elementary school to

be crucial to the children's attitudes that are in their formative stage. Out of a total of 27 experiments done in the area of frequency of informal achievement testing, only one study was done at the elementary school level.

Mann, Taylor, Proger, and Morrell (to be published) dealt with daily testing in third-grade arithmetic. The study material was multiplication. This pilot study was conducted in Spring, 1967, to determine whether or not being under the psychological threat of frequent testing in a natural learning situation during a unit of instruction will result in beneficial content structuring effects, increased attention to material, and so on, in terms of immediate and delayed retention.

Four randomized groups of about twenty students each were formed: BE, GE, BC, and GC (E, C, B, and G represent "experimental", "control", "boys", and "girls", respectively). To control for differing teacher effectiveness and possible interaction effects of teacher personality with students' personalities, the four teachers were randomly rotated throughout the four groups from day to day. Within each group, low and high previous achievement categories were identified (ex post facto) on the basis of the final arithmetic mark in second grade. The two E sections received practice worksheets which were to be counted in with their total mark and accordingly were letter graded, while C received the identical worksheets and were told they would count only as practice. All four groups received the worksheets back in class the

next day and were told to locate and correct the errors they had made. During most of the rest of the class period, the new worksheets were used. The experiment lasted twenty class days.

The reviewer is presently analyzing the results of this experiment. The design is totally confounded across blocks (groups) with respect to methods and sex. However, within each block, an unconfounded comparison can be made between high- and low-previous achievement subgroups. Although the confounding in this design was inadvertent, it actually aids the experimenters in studying a particular aspect of the problem: the interaction of methods by previous achievement, a confounded design in this specific case gives a more powerful test of such a measure. The investigators were especially interested in testing whether or not high previous achievers would do better in E than C (that is, good students might like the challenge of a test condition rather than a practice condition) and low previous achievers would do better in C than E (that is, poor students might feel more secure under the nonthreatening practice condition than the experimental one).

The study by Mann et al. (to be published) was the only one at the elementary school level on the topic of frequency of informal achievement testing. This is one reason why the reviewer decided to do his dissertation experiment on comparing daily testing, alternate-day testing, once-a-week testing, and no testing in sixth-grade arithmetic. With the Mann et al. (to be published) study above, the reviewer's dissertation, and the experiments

to be described below at the junior and senior high school and post-high school levels, developmental implications might arise.

(b) Junior and Senior High School: Of a total of seven experiments at both the junior and senior high school levels, only one study touched upon the junior high school level: Maloney and Ruch (1929). They compared three methods of teaching grammar in ninth, tenth, and eleventh grades in junior and senior high school. Three methods groups were formed from a total of 497 students. The first group used only the textboo' and was given no tests. In the second group, ten 25-item tests were used as instructional material in place of the textbook; another five 25-item tests were used for evaluative purposes. The third group was taught by a combination of the textbook and five short tests.

Although no significance of results was stated, a trend was noted with the test group achieving highest and the combination method next. The reader should note that, while lacking perfect control, the three methods have pedagogical soundness. The reviewer considers this experiment to be "broad curricular research"; a natural learning situation prevailed. However, one will never know just what it was about each method that caused the weak but notable trend in results.

No other studies on frequency of testing occurred at the junior high school level (grades seven through nine). Thus, the studies that dealt exclusively with the senior high school level

are considered next.

Kitch (1932) studied the effect of frequent informal achievement tests when used only as practice (not counted in with the students' grades). The students were enrolled in tenth-grade high school biology. C consisted of the group of 89 students who took the course the preceding year, while E was the group of 88 students currently available. No attempt at matching the two groups was made, since the investigator found that the difference in average Terman Intelligence Test scores for the two groups resulted in an insignificant Z value (P = .98); hence, he considered the two groups to be initially equal for his purposes.

Short, frequently-given practice tests were given to E but not to C. E was told that these short tests would not count in their grade. Unfortunately, nowhere in his exposition does Kitch say just how often such practice tests were given to E; one would assume that E was given at least one practice test for each unit. The first four units of instruction were covered in this manner. For the fifth unit of instruction, no practice tests were given to E; in effect, other than carry-over effects, E and C were treated alike. During the first four units, the next day the papers were returned to each student so that he did not receive his own paper; the papers were then corrected by the pupils.

Both E and C received the same major unit tests. These tests formed the main criteria of comparison. "Though there had never

been any attempt to standardize these tests it was believed that they were sufficiently valid and reliable for the purpose of this experiment. Observations made during the period in which these tests have been used indicate that with comparable groups, comparable scores have been made." (Kitch, 1932, p. 39). Using number of errors made, the investigator presents simple Z ratios for each unit: on the first unit, C > E ($P_{2\text{-tail}} = .02$); on the second unit, C > E ($P_{2\text{-tail}} = .18$); on the third unit, C > E ($P_{2\text{-tail}} < .01$); on the fourth unit, C > E ($P_{2\text{-tail}} = .20$); and on the fifth unit, C > E ($P_{2\text{-tail}} = .13$).

Kitch concludes that such self-scored practice tests are well-worthwhile in subject areas where many facts are presented; the teacher is saved a lot of time from paperwork, while at the same time the student is motivated to prepare himself better. The reader should note, however, that, as far as motivating students to prepare better outside of class is concerned, such would probably not be the case in elementary school where most students do not as yet take their school work very seriously.

Connor (1932) investigated the effect of frequent testing in high school physics. From a working pool of seven experimental classes and ten control classes, he formed four matched groups on the bases of mathematical ability (Kilzer-Kirby Inventory) and intelligence (Otis Self-Administering Test of Mental Ability). The experimental groups made use of the "Instructional Tests in Physics" by Glenn and Obourn. Twenty of these tests were given to the

experimental students during the school year. However, when measured on the two posttest criteria (Harvard Elementary Physics Test and the 1930 Iowa Academic Meet Test in Physics), the control groups, on the whole, exceeded the experimental groups. Connor attributes his results to the fact that the twenty frequent tests took away instructional time from the experimental groups; he does not advocate the use of such frequent testing in high school physics.

Connor's procedure can be termed "broad curricular research"; if a program of frequent testing is to be used, instructional time apparently must decrease in relation to the control group (such need not be the case, if administrative difficulties can be overcome, as was the case in the reviewer's dissertation experiment). An investigation such as Connor's study is useful for comparing broad curricular programs of frequent testing. However, "broad curricular research" will never answer the more psychological, noncurricular problem of whether or not the test itself, by virtue of its threatening, negative aspects or positive, motivating effects, can actually teach the child something just through increased attention to the matter at hand. This type of reasoning refers only to what occurs during the taking of the test in class, not to the motivating of the student to prepare for the test outside class. Both issues are important; however, the former, being more "basic" in nature (Psychologically oriented) and harder to control in research, has usually been ignored. Again, one should be able to see the distinction between "basic" research and "curricular."

McClymond (1932) conducted a study very similar to that of Connor (1932): using or not using the Glenn-Obourn practice tests in high school physics. However, McClymond did get significant achievement results in favor of the students who had used the Glenn-Obourn tests. But the investigator points out that the testing time was subtracted from the laboratory periods and not from the lecture periods, as in Connor's study; McClymond's procedure thus resulted in greater control than in Connor's experiment.

Weissman (1934) investigated daily testing in high school physics. E (181 students) received daily tests for about nine weeks; C (180 students) did not receive any daily tests. The groups were matched as far as possible on chronological age, cumulative mathematics average, cumulative physics average, and pretest score. It was found $(E - C) = 6.69 \pm 0.89$ $(P \ll .001)$.

Curo (1963) studied the effect of daily quizzes on eleventh-grade American History classes. Three types of schools were used: large metropolitan, medium suburban, and medium rural-consolidated. Ten intact classes were divided into five experimental groups and five control groups. "To minimize the effect of varying teacher competence, each teacher instructed one or more pairs of control-experimental classes." (Curo, 1963, p. 70). The experiment lasted six weeks.

A cumulative pretest was given to all E and C classes; the pretest was the First Semester American History Test, State High

School Testing Service for Indiana. Otis Mental Measurement scores were also available. The control classes received two major unit tests but no daily quizzes during the six-week study; the experimental classes received the daily quizzes but not the major unit tests. (Because of this inequity in the amount and kind of testing, the reviewer classifies Curo's study as "broad curricular research".) For the E classes, the daily quiz was given during the first five to ten minutes of the class period on a previous study assignment. The investigator devised his own 135-item posttest. Also, all classes were given the original pretest again as a second immediate posttest.

In the analysis of variance for the original pretest, the two classes of the suburban school were so discrepant in variance (97.63 versus 85.78, although no test of significance is provided) that the whole school was deleted from all subsequent analyses. Then, eight intact classes were left for a total of 184 students. To achieve equal cell frequencies of 46 (sum of all students within one school for one method), random selection was used. On the original pretest, no significant differences were found for methods, schools, or methods by schools (all F's < 0.50). When the same pretest was used as a posttest, schools came closest to significance ($P \doteq .30$), but again methods and methods by schools were definitely insignificant (both F's < 1.00). On Curo's own posttest, schools were significant ($.025 < P < .05$), methods were next ($.10 < P < .25$), and methods by schools last ($P \doteq .30$).

Curo concludes that using daily informal achievement tests to increase the learning of factual material is not effective at all. Perhaps Curo's analyses would have yielded more accurate conclusions if he had used analysis of covariance instead of his three separate analyses of ordinary variance. Further, his inspection method of deciding whether or not to delete a school or pair of classes on the basis of apparent nonhomogeneity of variance is open to criticism; rather, a transformation of scores is suggested.

The last study at the senior high school level on frequency of testing is Pikunas and Mazzota (1965). They studied the effects of weekly testing in twelfth-grade chemistry in a large city technical high school. A total of 128 students were taken from four intact classes. Two intact classes were assigned at random to the first treatment group, while the second treatment group received the other two intact classes. Each class met for three 68-minute periods a week. The study lasted twelve weeks. One chapter per week was covered. The twelve chapters were divided into two independent sets of six each: A and B.

A crossover design was used. During the first six weeks, both groups of classes did not receive weekly tests; the first treatment group (two intact classes) had the A chapters, and the second treatment group (two intact classes) had the B chapters. The six-week examination was from the publisher of the textbook chapters. This test was not returned to the students after they had taken it. All tests in the study (the six-week criterion and the weekly noncriteria)

were corrected by a teacher not involved in the instruction of the experiment. During the second six weeks, both groups of classes received a quiz once a week, again taken from the publisher's test booklet for his textbook. However, this time the first treatment group received B, and the second treatment group received A. The lost teaching time because of the weekly tests during the second six weeks was compensated for by giving enforced study periods during the first six weeks.

The analysis supplied by the investigators consists only of crude percentage statistics pooled across chapter sets and treatment groups to obtain E versus C ($\bar{E}$ = 70.84% and $\bar{C}$ = 60.77%). No statistical test was made. The investigators caution, "There is, of course, a danger of becoming too preoccupied with testing, and of allowing this preoccupation to lead to a distorted situation where testing is credited with attributes and accomplishments which it does not possess." (Pikunas and Mazzota, 1965, p. 375). They also claim, "Additional investigations are necessary to find out whether this also applies to other subjects than science and whether this applies on each level of education." (Pikunas and Mazzota, 1965, p. 376).

In the above experiment, to form the comparison of major importance--E versus C--one must admit the flaw of "history" and, to a lesser extent, "maturation" (as defined by Campbell and Stanley, 1963). The reviewer cannot see the logic of this particular cross-over design. Apparently the use of two different sets of chapters

at the same time was to counter any exchange of information from one group to the other, but this is of no benefit to the E versus C comparison in the present design anyhow. Further, with the original design, one cannot determine whether the apparently superior performance of the E groups of classes (that is, all the classes during the second six weeks) was caused by the weekly tests or by the perhaps better-activated study habits of the students, having already gone through six chapters of the same publisher's format.

(c) Post-High School Experiments: Nineteen studies fall into the post-high school category. Before reviewing the experiments, however, the reader should note that any attempt to generalize the findings of college studies down to the high school and, especially, the elementary school might be on very shaky ground. The college milieu is very different from that of the public schools. College classes meet only two or three times a week and sometimes only once a week; the students are allowed much more responsibility than in their pre-college days; they have much more free time to do with as they please. In fact, within the public school system, it is true that even high school and junior high school are psychologically different. Such psychological differences among educational levels become of crucial concern in the elementary school where the child, still in his formative years, first meets the threat of informal achievement tests. It is here that least is known about the way informal tests can be used as learning devices, and it is here that generalization from the college situation is

most dangerous. For one thing, students cannot be expected to prepare very extensively, if at all, for tests on their own time outside class; rather, one must concentrate attention on the problem of what effect informal tests have on his in-class performance, assuming most, if not all, preparation for tests must occur in class itself.

Deputy (1929) studied frequent testing in second-semester freshman philosophy at a state university. The testing material consisted of the preceding day's lesson. Each class met two days a week. At the start of the semester all students were given the Otis Self-Administering Test of Mental Ability: Higher Examination. At the time of the experiment the reliability of the Otis test was given as .92. Two experiments were performed on the same students of three intact sections: one study during the first half of the semester and one study during the second half of the semester.

During the first half of the semester, three different testing methods were compared. Section E_1 (30 students) received a ten-minute quiz every class day. Section E_2 (33 students) received a twenty-minute quiz once a week. Section E_3 (33 students) received no quizzes or unit tests. The investigator paid especial attention to rationalizing to the classes the differences that existed among their procedures. Code numbers were given to each student so that only he would be able to recognize his results on the blackboard the next day. This first experiment lasted for six weeks until the mid-semester test.

During the first half of the semester, the z test results for the Otis test were: $E_1 > E_3$ ($P_{2\text{-tail}} = .07$); $E_2 > E_3$ ($P_{2\text{-tail}} = .46$); and $E_1 > E_2$ ($P_{2\text{-tail}} = .20$). Since none of these initial difference measures was statistically significant, Deputy goes on to discuss differences on the criterion posttest (mid-semester examination). On this test significant z test results were found: $E_1 > E_3$ ($P_{2\text{-tail}} \ll .001$); $E_3 > E_2$ ($P_{2\text{-tail}} = .13$); and $E_1 > E_2$ ($P_{2\text{-tail}} \ll .001$). In summary, E_1 significantly outperformed both E_2 and E_3, while E_2 apparently was not even as effective as E_3.

During the second half of the semester, E_2 received a ten-minute quiz each class meeting, while both E_1 and E_3 received no quizzes at all. On the final examination (about 100 items), insignificant z test results were obtained: $E_1 > E_3$ ($P_{2\text{-tail}} = .20$); $E_2 > E_3$ ($P_{2\text{-tail}} = .05$); and $E_2 > E_1$ ($P_{2\text{-tail}} = .27$).

The analyses of both Deputy experiments can be criticized on the grounds that simple z tests were used instead of a combination of analysis of variance and multiple comparisons. In fact, the pretest Otis scores should have enabled an analysis of covariance. Further, the final examination performance of all three groups during the second half of the semester is contaminated by unequal carryover effects from the first half of the semester.

Serenius (1930) dealt with frequency of testing in college history. The looseness of the design would qualify it as "broad curricular research." Two classes (totaling 64 students) each

had the same lecture class twice a week. During the third meeting of each week, E received a test, while C had informal discussion. At the end of the semester no significant differences were found.

Turney (1931) investigated frequent testing in an educational psychology course with college juniors and seniors. Two intact groups were used. The final examination criterion posttest was divided into two 175-point equivalent forms A and B. Both forms were a combination of true-false, multiple-choice, and completion items. Form A was given as a pretest to both groups at the start of the study. The group scoring lowest on this pretest became E (N_E = 40 and N_C = 28). The final examination consisted of both forms A and B. A different mid-semester examination (neither A nor B) was given to both groups. E received eleven short quizzes throughout the semester about once a week; C had only one quiz. The quiz results were given to the E students at the next class meeting but the quizzes themselves were not returned. Both E and C were taught by the same instructor.

Using a third intact section of students comparable to those in the study, Turney claims no practice effects could be found between forms A and B and that the difficulties of the two forms were equal. Because of the results obtained from the third group not directly involved in this study, for computing gain scores, Turney subtracted two times the pretest form A from the final test (both A and B). Because of the way he chose group E, on the pretest the z test found C $>$ E ($P_{2\text{-tail}} \doteq .0001$). On the final examination, however, E $>$ C

($P_{2\text{-tail}} \doteq .99$). For gain scores themselves, E > C ($P_{2\text{-tail}} \doteq .005$).

Smeltzer (1931) studied both frequent testing and exemption from certain class work (by test performance) in a course in undergraduate educational psychology. The course was given repetitively in a university on the three-term schedule. Throughout each of the three terms an attempt was made to standardize presentation of material as much as possible by making both E and C classes follow the same course outline for the seven major topics of the course; a calendar of topic progression was also used. In each term, the procedure for the C intact classes was the same. C classes received two fifty-minute essay tests during the term. In each term, all E and C classes were given a pretest and an objective final examination. Any test given to either the E or C classes was scored by the teachers and given back to the students at a later class meeting; the tests were corrected so that not only were incorrect answers marked wrong but also the correct answers were provided. In both E and C classes the discussion-recitation method was used. The pretest and final examination were parallel forms of the same objective examination. Throughout the three terms most of the total of 523 students were freshmen and sophomores.

During the autumn term six intact classes (three E and three C) were used. In the E classes, a major objective examination was given every other Thursday outside of regular class in the late afternoon. Then, on the immediately following Friday morning, the corrected tests along with an item analysis were ready for the E instructors to use

during the regular class meeting that day. During the first twenty minutes of this Friday morning class, the E students received their corrected tests and a discussion of the test questions was given. "At the end of the [E] class period on Friday each instructor would read the names of approximately one-fourth to one-third of his students who scored highest on the examination and who would therefore be excused from class the following Monday and Tuesday." (Smeltzer, 1931, p. 31). The students who were required to return on the next Monday and Tuesday underwent an intensive discussion of the subject matter on the E test of the preceding Thursday; these low-scoring students then were given a twenty-minute retest on this subject matter. On Wednesday, both low- and high-scoring E students resumed their usual classroom procedures.

The analysis of percentiles on the final examination during each term was carried two times: (1) without matching and (2) with matching. (However, the reviewer sees no advantage to examining unmatched results when one has the matched results also available). "A very rigid pairing involving three criteria was used. The criteria were on the basis of (1) sex; (2) intelligence test percentile to the extent of $\pm$ 4 points; (3) pre-test score to the extent of $\pm$ 5 points." (Smeltzer, 1931, p. 90). Five percentiles were compared on the final examination during each of the three terms: 90%-ile, Q_3, Median, Q_1, and 10%-ile. Simple z tests were then computed. During the autumn term only the 10%-ile (E-C) difference had a reasonably large z value: $P_{2\text{-tail}} \doteq .27$ (35 matched pairs).

However, it should be noted that, with the exception of the 90%-ile level, all comparisons were in the direction of E > C. One should note that the autumn experiment was of a confounded design (actually, a confounded, confounded design, if such can be the case; that is, confounding occurred not only across methods blocks but also within the methods blocks for different frequencies): two levels of frequent testing for E, by exemption.

During the winter term three intact E classes and three intact C classes were again used. However, this time the E classes received weekly twenty-minute tests every Thursday instead of the fortnightly tests. After the instructors corrected the E tests outside class, they were returned to the E students on the following Friday morning for about twenty minutes. As in the autumn term, the E students had to return the tests to the teacher at the end of this time. The same exemption procedure was used for determining which students had to return on the next Monday. However, the retest for the nonexempted students was given the latter part of Monday's class meeting after intensive review of the Thursday quiz's subject matter. Further, in contrast to the autumn term, each E student, regardless of being exempted or not, was only given numerical grades--no letter grades. Also, using the Thursday quiz score, on the average of the quizzes of Thursday and Monday, a graph of each student's weekly progress was kept in the classroom on a large chart.

During the winter term, only the Q_1 and 10%-ile (E-C) comparisons provided z tests that are worth noting: $P_{2\text{-tail}} \doteq .37$ and

$P \doteq .004$, respectively. However, at each of the other three percentile levels, $E > C$. Again, one has a confounding within confounding design: frequency within frequency by exemption by motivation.

During the spring term, E group still received the same type of weekly testing. However, the exemption process was modified so that anyone could omit the Monday class, or attend it, as he wished. Further, the E students were now given a weekly chart-in-class incentive on average letter grade (Thursday alone or Thursday averaged with Monday, as the case might be), as compared to the winter term's numerical chart value incentive. There were three intact E classes and four C classes. The C procedure remained unchanged.

During the spring term, again only the Q_1 and 10%-ile (E-C) comparisons provided z tests worth noting: $P \doteq .06$ and $P \doteq .01$, respectively. The same design criticism can be made here as with the winter term: confounding within confounding (frequency within frequency by exemption by motivation).

Kulp (1933) studied the effect of weekly quizzes on graduate students in educational sociology. Two one-hour class meetings per week were scheduled. A ten-minute quiz was given each week to all 32 students until the mid-semester examination. "On the basis of the mid-term examination, the class was divided into a 'high' half and a 'low' half. . . . Following the mid-term examination, only the 'low' half of the class took the ten-minute weekly tests;

the 'high' half was excused." (Kulp, 1933, p. 158). Using a z ratio to compare the difference between "halves" on the mid-semester examination, H > L ($P_{2\text{-tail}} < .01$); however, on the final examination, H > L ($P_{2\text{-tail}} = .36$). Because of the large loss of significance of the difference between H and L during the second half of the semester, Kulp concludes that the weekly tests benefited L "significantly."

Keys (1934a) dealt with weekly testing in educational psychology during the spring semester. Although 360 students comprised the experimental and control sections, matching on the basis of sex and a 167-item true-false pretest reduced the total to 143 analyzable students in each section. Keys divided the semester into three equal parts of four weeks each. Although his main interest was in weekly tests versus monthly tests, he confounded this comparison with the use of study guide assignments in E. The testing procedure itself was carefully controlled :". . . tests administered to the two sections were identical, both in content and total amount, differing only in that the experimental group took these in brief weekly installments, and the control in the form of long mid-term examinations." (Keys, 1934a, p. 428).

During the first four-week period, E had specific weekly assignments and four weekly tests, while C had a general, monthly assignment and one monthly examination. During the second four-week period, E again had specific weekly assignments and four weekly tests, while C also had specific weekly assignments but

only one monthly examination. During the last four-week period, E had only the general monthly assignment and one monthly examination, while C had specific weekly assignments and one monthly examination.

To make comparisons between testing modes for each four-week period, Keys added up the four weekly tests of E to pit the sum against C's one monthly test. For the first four-week session, $E > C$ ($P_{2\text{-tail}} < .001$); for the second four-week session, $E > C$ ($P_{2\text{-tail}} < .001$); for the third four-week session, $E > C$ ($P_{2\text{-tail}} \doteq .03$); and for the final examination, $E > C$ ($P_{2\text{-tail}} \doteq .13$). Thus, while differences exist between groups during the study, the groups are not much different when cumulative achievement is measured at the end of the study.

Considering equality of study-guide assignments in Keys' study, only during the second four-week period can a reasonable comparison be made between weekly tests and monthly tests. However, such a comparison would be contaminated by any carryover effects from the first four-week period. In fact, the first four-week period is the only one that allows any comparison free of contamination; however, no matter what comparison is made at any time throughout the study, there will always be confounding present. This is an extremely complicated design situation of which the analytical considerations Keys was unaware.

Eurich, Longstaff, and Wilder (1937) investigated a program of weekly tests in an introductory college psychology course. The course

met three times a week with one hour and twenty minutes to each class. The experimental group consisted of thirty-eight men and twenty-five women drawn from a total of 138 students, while the control group was chosen on a matching basis (College Ability Test and course pretest) from the previous fall semester's 288 students. Both C and E were given the same 364-it course pretest at the start of their respective semesters. pretest covered (1) facts and principles, and (2) attitudes and beliefs. E was marked in terms of the percent of possible gain left after taking into account the pretest score. On the other hand, C was graded the usual way on the basis of the mid-semester and final examinations. E was given the original pretest as its final examination, while C had a composite of the first five weekly tests of E as C's mid-semester examination and a composite of the last five weekly tests of E as C's final examination. Then to get a total score for C that would be comparable to E's final test, C's mid-semester and final examination scores were added.

The investigators used the odd-even method of determining the reliability of their tests. "The reliability coefficients thus derived for the experimental group are: initial test, .95; sum of weekly tests, .97; final test, .97. The reliability coefficients for the control group are likewise high: initial test, 194; final test, .95." (Eurich et al., 1937, p. 336). Using simple z ratios, the investigators found that for the College Ability Test, E > C ($P_{2\text{-tail}} = .13$); for the pretest, E > C ($P_{2\text{-tail}} = .32$); and

for the final test, E > C ($P_{2\text{-tail}}$ = .49). Thus, as far as "facts and principles" are concerned, Eurich et al. conclude that the combination of weekly examinations and grading on the basis of relative potential gain was not effective in increasing achievement. Further, using the extremes of the distributions (upper and lower sevenths), the same insignificant results were obtained. When the investigators broke their posttest up into the ten major topics covered throughout the course, technical material appeared to benefit more than the more generalized areas.

The procedure of matching spring semester students with the preceding fall semester's students can be criticized on the basis of "history": extraneous events that occur during one semester and not the other. Further, the psychological atmosphere of the spring semester is different from that of the fall semester; the drudgery of the beginning school year and the major vacation of Christmas affect the fall semester, while the results of the spring semester are affected by the students' carefree attitudes. Also, frequency was confounded with grading procedure. The procedure used with E can be criticized on the basis of pretest sensitization; a parallel form of the pretest would have been more appropriate. Moreover, E could be said to have undergone practice effects with respect to the final examination, since the ten weekly tests were made up directly from the items of the lengthy pretest.

Johnson (1938) found that weekly tests resulted in significantly higher achievement on an immediate posttest for college

students. However, this difference vanished on a delayed posttest six weeks later.

Noll (1939) studied giving frequent quizzes in an educational psychology course for third-and fourth-year students at the university level. During one year, an intact class was given four quizzes at approximate intervals of three weeks each (E). During the following year, another intact class was not given the quizzes (C). Both classes were compared on the 100-item, one-hour objective mid-semester examination and the 200-item, two-hour objective final examination. The quizzes were graded by the instructor and returned later for discussion. The students of both groups were matched on cumulative university grade point average and American Council Psychological Test score.

Thirty-three matched pairs of students were analyzed by z tests. On ACPT, a z test showed the two groups to be comparable ($P_{2\text{-tail}} \doteq .90$); similarly, the GPA means were comparable ($P_{2\text{-tail}} \doteq .91$). On the mid-semester examination, C $>$ E ($P_{2\text{-tail}} \doteq .60$); on the final examination, C $>$ E ($P_{2\text{-tail}} \doteq .41$). Similar insignificant differences were obtained when the eleven highest and eleven lowest students in both groups were analyzed separately. "There is no evidence here, and little in other studies, to support the common belief among instructors that written tests as commonly used motivate learning or increase total achievement in college classes. If they do add something to the effectiveness of our teaching, this fact remains to be demonstrated with other measures than those used

in studies reported to the present time." (Noll, 1939, p. 357). The reader should take note, however, that the feedback mode has been confounded with the frequency mode.

Ross and Henry (1939) conducted a study in a general psychology course at the university level. Both E and C received the same pre-test, mid-semester examination, and final examination. E also was given weekly objective tests. On the final examination, E achieved significantly higher than C. However, in an identical study in educational psychology, opposite significant results were obtained.

Sumner and Brooker (1944) dealt with daily testing in a general psychology course at the college level. The course records of 200 students were analyzed. After two or three weeks from the start of the semester, forty daily tests of matching type were given to the class; about 25 items were on each daily test. No control group was used; the investigators were interested only in the predictive value of the average of the first five quizzes in relation to the average for all forty quizzes. For the whole group, $r_{xy} = .82 \pm .0156$; the z test for the difference in average percentages of the first five quizzes and all forty quizzes is statistically significant: $P_{2\text{-tail}} \doteq .002$. The z test, however, appears to be invalid; related samples can be compared, but the measures themselves must be taken independently of one another. Although the same criticism could be leveled at the r_{xy} calculation, the purpose for which it was computed (prediction) seems to make the method valid here. The investigators conclude, "The standing of the students relative to one

another at the end of the first 5 tests will be approximately the same as their standing relative to one another at the end of the 40 tests." (Sumner and Brooker, 1944, pp. 323-324).

As a further analysis, the students were divided (on paper) into a "Hi-Lo" dichotomy on the basis of their averages for the first five daily tests; the "Hi" category consisted of the 114 students with a five-test average of 70% or above, while the other 86 students formed the "Lo" group. The same type of Pearson product-moment correlation coefficient as for the entire group above was computed separately for the "Lo" and "Hi" groups: $r_{Hi} = .60$ and $r_{Lo} = .68$. Hence, the authors conclude that the students having an initially low average gain much more than their high-standing counterparts throughout the semester .

Fitch, Drucker, and Norton (1951) studied the effect of weekly quizzes on achievement in an introductory college government course. The intact control class of 97 students had only the usual monthly quizzes, while the intact experimental class of 198 students had not only the usual monthly quizzes but also the weekly ones. Both classes met three days a week with the same professor and both sections had access to voluntary discussion sections outside of class. During the third class meeting every week, the last half hour for both classes was set aside for questions and discussion; however, ten minutes were taken away from this time for E to have its weekly quiz. Four one-hour monthly tests were given to both groups. The weekly quizzes that E received were on textbook assignments,

not on the lecture material. The criterion variable was the grades of the five hour-long tests given to both groups (Y). Because of the intact classes, the covariate (X) was the preceding semester's government course grade (the government course was a two-course sequence). The regression of Y on X was found to be linear.

For purposes of analysis, the degree of voluntary discussion group attendance was broken down into four frequency classes. A two-way analysis of covariance was used; methods (two levels) by degree of discussion group attendance (four levels). To achieve equal total numbers for E and C, 91 students were randomly selected from the original E pool so as to use all of the 91 C students for whom complete discussion attendance records were available. However, within the total equalized E and C groups, the corresponding frequencies for a certain level of discussion frequency were disproportionate; the analysis of covariance was adjusted accordingly. For the total sample, Y was found to be almost normally distributed. Homogeneity of residual variance was found to be satisfied by Hartley's M-test. Using the method of disproportionate cell frequencies (nonorthogonal case), Fitch et al. found $E > C$ $(.01 < P < .02)$, monotonically increasing achievement in relation to increasing voluntary discussion session attendance $(.05 < P < .10)$, and methods by discussion interaction insignificant $(.50 < P < .75)$. The investigators conclude, then, that regardless of discussion group attendance, weekly quizzes significantly aid achievement in college government courses.

One should note that the main effect (weekly quizzes) is confounded with other variables in the experimental group (grading practices). The reviewer believes that confounding of frequency of testing with grading procedures exists because the students were told that the quizzes would not be counted in their final grades. The question raised by the reviewer is how one can have a test without a grade. Further, for reasons which elude the reviewer, the investigators first proceed with separate analyses of covariance for methods and for discussion group attendance. Finally, they present the two-way analysis of covariance (methods and discussion group attendance). What the two one-way analyses of covariance were supposed to accomplish, the reviewer does not know. Without the measurement of the interaction between methods and discussion group attendance, any discussion of either main effect by itself is suspect. Hence, only the two-way analysis should have been used to begin with.

Guetzkow, Kelly, and McKeachie (1954)--doing curricular research rather than more basic research--dealt with three methods of teaching an elementary general psychology course in a university. Twenty-four intact classes were used (N = 25 to 35 for a single class). For the first class meeting of each week, the twenty-four classes were pooled into three large lecture sections (N = 250 to 300 for a lecture section); each large lecture section received the same one-hour lecture. For the other two one-hour class meetings, the students reverted to their original grouping among the 24 classes.

Eight teaching fellows taught three classes each; each fellow had to use all three methods. However, each class was exposed to only one method. It was during the small-class meetings (last two sessions every week) that the experimental manipulations took place. In the first method (M_1), a short completion quiz was given at each of the two small-class meetings; the instructor was told to lead and structure all discussions. In the second method (M_2), a discussion atmosphere was created; the students structured the meetings. Other than course examinations, essay tests were given in M_2 once every other week. In the third method (M_3), independent study was stressed; the instructors were there only for consultation. No quizzes were given. Extra reference material was provided for making the independent study more flexible.

The reader can see why the reviewer considers this type of experiment as "broad curricular research." The design of the experiment was dictated by purely on-the-job considerations without regard to confounding. If any differences arise, one will never know what particular aspect of the superior method caused it to be so; the results have to be taken at face value, realizing their limitations.

Each of the eight instructors had responsibility for each of the three methods; thus, hopefully, each instructor would not be prone to favor one method over the other just from sole usage of that method. "The time sequence of methods was varied from instructor to instructor so that no method would consistently be

used for the instructors' first sections to be taught during the week. All permutations of the order of methods were used." (Guetzkow et al., 1954, p. 198) Students were changed to different groups at the start of the experiment so that each class". . . had an appreciable number of women, veterans, etc., and mean score intelligence and grade point average were perfectly matched." (Guetzkow et al., 1954, p. 199) The students were given a rationalization as to why the different methods were being used.

Three major examinations were given to all classes throughout the course: a major unit test (covering the first four weeks), the mid-semester examination (after eight weeks), and (T_1) the final examination. Other posttest criteria were also used: (T_2) the United States Army Forces Institute Examination in elementary psychology, (T_3) a test made by the eight instructors of common misconceptions in psychology (corrected split-half reliability of .73), (T_4) Duvall's Conceptions of Parenthood Tests (corrected split-half reliability of .80), (T_5) McCandless' Scientific and Analytic Attitude toward Human Behavior, (T_6) an attitude-toward-psychology test made by the eight instructors (corrected split-half reliability of .75), (T_7) the number of students planning to concentrate in psychology, and (T_8) the number of advanced psychology courses students want to take.

Although separate one-way analyses of variance (better yet, analyses of multiple covariance) would have been more appropriate for this three-method study, the investigators supply simple t tests

of the difference between the highest and lowest means. No significant differences were found for T_2, T_3, T_4, or T_5. One would assume from the investigators' description that for each pair of methods being compared, $df_t = N_1 + N_2 - 2 = 14$ (using intact classes as the unit of analysis). Then, on T_1, $M_1 > M_3$ ($.01 << P_{2\text{-tail}} < .02$); on T_6, $M_2 > M_3$ ($.05 << P_{2\text{-tail}} < .10$); on T_7, $M_1 > M_2$ ($.02 < P_{2\text{-tail}} < .05$); and on T_8, $M_1 > M_3$ ($.01 < P_{2\text{-tail}} << .02$). In spite of the deliberate attempt of the investigators to make the three methods as different as possible, Guetzkow et al. conclude that the methods are not impressively different.

The documentation of analyses is very poor. Although only t tests of the differences between methods are provided, the investigators make a statement that implies that they had made some kind of analysis of variance or covariance: "Not only are the differences between instructors not statistically significant but there was no significant interaction between instructor and method." (Guetzkow et al., 1954, p. 202). However, data is not provided.

Maize (1954) investigated two methods of teaching composition in remedial English classes at the university level. Themes written in class were treated as quizzes. One group wrote forty themes in class throughout the semester; each day's themes were criticized and discussed in class. A second group used a combination of workbook drill (English usage) and the writing of fourteen themes; however, the themes were corrected outside class by the instructor.

The first group achieved significantly higher on a test of English usage. However, here it must be noted that method of correction and method of feedback were totally confounded with frequency of quizzing (writing the themes). This situation need not have been the case if proper design had been exerted. No doubt the inequity of time allotment with respect to the use of workbook drill could not be avoided; the testing time has to be gotten somewhere. However, there is no pedogogical logic or necessity for confounding the results with modes of correction and feedback.

Mudgett (1956) studied the effect of daily, weekly, and monthly testing on achievement in engineering drawing. The course was meant for first-year technical students (engineering, mining, metallurgy, mathematics, and so on). Eight intact classes were used in this experiment; they were randomly chosen from a total of 21 sections of the course. Then the selected classes were assigned at random to three different testing programs. Randomization had also been used at the registration period, where students had been assigned to the original 21 classes at random. Further, the four instructors used in this experiment were assigned to the sections at random. All groups had testing time taken out of the laboratory drawing time, not from the lecture periods. Each group met for eight periods a week: two hours for lectures and six hours for laboratory drawing work. For purposes of analysis, 184 students formed the final sample. The design appears to be a type of balanced, incomplete-block design:

INSTRUCTORS

TIME	A	B	C	D
8:30	DAILY TESTS	MONTHLY TESTS	MONTHLY TESTS	WEEKLY TESTS
10:30	MONTHLY TESTS	DAILY TESTS	WEEKLY TESTS	MONTHLY TESTS

(Mudgett, 1956, p. 58). Attempts to keep teaching procedures relatively uniform were made by having all instructors follow a course outline and attend weekly meetings with the investigator.

Two classes (E_1) received a ten-minute quiz at the start of each class period. The students corrected the papers in class that same day, and the results were discussed at that time. Two classes (E_2) received a thirty-minute test at the end of each week. The tests were corrected by the instructors over the weekend and returned the following class meeting, where discussion of the tests was provided. Four classes (E_3) received a major unit test at the end of the fourth and ninth weeks of the semester, as well as the final examination. No other tests were given. "These unit tests were machine scored and only the scores were given to the students; hence, the four classes in the Monthly Test Group knew their class standings but had no other specific information to be used to adjust study techniques." (Mudgett, 1956, p. 4).

More specifically, E_1 (two classes) received 34 ten-minute quizzes; E_2 (two classes) received 8 thirty-minute tests. E_3

(four classes) had only the four major tests. The four major tests formed the criterion measures for this experiment: Engineering Drawing Test Form A (T_1); Engineering Drawing Test Form B (T_2); Engineering Drawing Performance Test (T_3); and Engineering Drawing Theory Test (T_4). T_1 was given as a pretest and again at the end of eight weeks. T_2 was given at the close of the first four weeks of instruction. T_3 and T_4 were given as immediate posttests. The 50-item, multiple-choice T_1 had reliabilities of .86, .81, and .80 for three different samples when Hoyt's analysis of variance technique was used. The 40-item, multiple-choice T_2 had a corrected Spearman-Brown reliability of .84 and a Hoyt reliability of .81. The 50-item, multiple-choice T_3 had a Hoyt reliability of .65. The 200-point, combination-type T_4 had a Kuder-Richardson 21 coefficient of .96.

Analysis of covariance was used for evaluating the criterion posttests of T_1, T_2, T_3, and T_4; T_1 (as a pretest was used as the covariate for all criterion tests, and T_2 was also used as a separate covariate in the case of T_4. The covariates were chosen on the basis of their significant correlation with the posttest criterion in question. For each of the five covariance analyses (two for T_4), the Welch-Nayer L_1 test for homogeneity of residual variance was run, as well as the usual analysis of variance test for homogeneity of regression coefficients. All covariance analyses were run with the usual between groups and within groups breakdown (seven df for between groups and 175 df for within groups). On T_2, the

adjusted means were insignificantly different ($.05 < P << .10$).

On T_1 used as a posttest, the adjusted means were significantly different ($P << .005$). Hence, for T_1, multiple comparisons were used next (one df for between groups and 175 df for within groups). $E_1 > E_3$ ($.10 < P < .25$). $E_2 > E_3$ ($.25 < P$). However, the comparison of E_1 versus E_2 was not given. Using a more detailed factorial analysis of covariance, Mudgett states that the interaction of instructors by methods was not significant at the .05 level. The original significance of the overall F ratio was caused by the significant comparison of instructors (A+B) versus instructors (C+D).

On T_3, the adjusted means are significantly different ($P << .005$). Therefore, multiple comparisons are again used. $E_1 > E_3$ ($P << .005$). $E_2 > E_3$ ($.10 < P < .25$). However, the comparison of E_1 versus E_2 was not given. The interaction of methods by instructors was not significant at the .05 level.

On T_4 (still using the T_1 pretest as the covariate, as in the analyses of covariance of T_1, T_2, and T_3), the overall adjusted means are significantly different ($.01 < P < .025$). However, this significance was caused by the significant comparison between instructors (A+B) versus instructors (C+D), and not by any multiple comparisons between methods. On the other hand, using high school rank as the covariate instead of pretest T_1, the overall adjusted means are again significantly different ($.025 < P < .05$). However, the investigator does not pursue this analysis any further. He concludes,

". . . there is no evidence to support the belief that students in engineering drawing who are given tests similar to those used in this investigation as frequently as once a week will learn any more effectively than the students who are given tests once a month." (Mudgett, 1956, p. 166).

Fattu (1957) reports briefly on a program of frequent informal achievement testing in an elementary engineering course for Navy enlisted men. The tests were directed more toward performance skills in the shop than toward theoretical learning in the classroom. However, the frequent performance tests were directly related to the engineering theory the men had learned in the classroom. The investigator compared the final classroom examination scores of classes that had passed through the program of frequent performance testing and earlier classes that had not. Although he does not provide tests of significance, Fattu presents the final classroom test means of two schools in which the improved testing programs were introduced: $E_1 - C_1 = 35.2$ and $E_2 - C_2 = 40.2$. It appears, then, that frequent testing in areas related to, but not a direct part of, classroom learning can increase classroom achievement.

Standlee and Popham (1960) studied frequent testing in an introductory educational psychology course for undergraduates. Four intact sections with a total of 104 students were used. All sections were taught by the same instructor. Section A had weekly quizzes of a twenty-item, true-false type; the instructor-corrected quizzes were

counted toward the students' final mark and were returned the next class period. Section B had the same weekly quizzes; the students corrected their own papers, and the grades did not count in the final mark. Section C had the same quiz material presented to them in a reading fashion by the instructor; he then answered the questions verbally himself. Section D had no quizzes in any form. For each section, the investigators postulated theoretical bases in various combinations of extrinsic motivation, knowledge of results, psychological structuring, and enforced activity with test subject matter. Sections A, B, and C each received a total of thirteen quizzes throughout the semester.

A 100-item, multiple-choice pretest was given to all four sections at the start of the study.

> At midsemester, a different 100-item multiple-choice type examination was administered to all subjects. Fifty of the test items were common to the pretest; 50 items were new. At the end of the semester, a 150-item multiple-choice type examination was given to all subjects. The test items included the other 50 items of the pretest, the 50 new items of the midsemester examination, and 50 new items. (Standlee and Popham, 1960, p. 323).

Analysis of covariance with the pretest scores as covariate was used. On the mid-semester examination, the adjusted means of the four groups were significantly different only in a marginal sense: Standlee and Popham claim $P < .05$; however, in fact, $.05 < P < .10$. Then, using multiple comparisons between specific adjusted means, only (A-D) was significant: $.02 < P < .05$.

On the final examination, the adjusted means of the four groups were not significantly different: $.10 < P < .25$. The investigators admit their design is confounded: frequency by correction by grading. The investigators conclude, ". . . the use of quizzes will tend to increase students' achievement of subject matter early in a lecture-discussion type of course, . . . but . . . the significance of the increase in achievement is lost by the end of the course. . . ." (Standlee and Popham, 1960, pp. 324-325).

Selakovich (1962) dealt with frequent testing in an introductory college course in American Government. Two classes were randomly created so that 19 students were in each. The students were then matched on the Cooperative American Government Test (Form Y). Both E and C were given three instructor-made, hour-long examinations during the course and Form X of the Cooperative American Government Test as a final examination. In addition, E was given twelve unannounced quizzes throughout the semester. Hence, one will not be able to determine whether or not any significant or insignificant differences between E and C are caused by frequent testing or by the effect of nonannouncement; the design is confounded.

The investigator used a t test for related samples. For the difference in cumulative means for the three major hour tests, $C > E$ ($P_{2\text{-tail}} > .20$). For the difference in means on the standardized posttest, $E > C$ ($.10 < P_{2\text{-tail}} < .20$). This is a strange reversal of results. Further, comparing the parallel-form, pre- and post-tests, $C_2 - C_1 = 3.32$ and $E_2 - E_1 = 3.84$; apparently, as

measured by the standardized tests, the students were not affected very much by the intervening course. It would be interesting to see whether or not, using Form Y of the CAGT as a covariate, the results for hourly tests and for the standardized posttest would be much different if an analysis of covariance is used.

Laidlaw (1963) studied weekly and monthly testing procedures during two semesters of a general psychology course. Each of three instructors taught a pair of classes; which of each pair of classes was to be the E group was randomly determined. Each of the six classes met for three one-hour sessions each week for 16 weeks a semester. Every third class hour, a 15-minute, 20-item, multiple-choice test was given to the three E groups. Results of these quizzes were made known at the next meeting but no discussion of questions was permitted. The three C groups were tested every fourth week with a one-hour, 80-item, multiple-choice test. The same feedback procedures as with E were used. Each of these monthly tests had 60 items in common with the equivalent weekly tests. The experiment was continued during the second semester to obtain measures of delayed achievement. Out of a total of 151 students, only 120 became the pool for analysis because of incomplete pre-experiment records. The three E groups had an original total of 87 students (but only 69 workable records), while the three C groups had an original total of 64 students (but only 51 workable records). In going from the first semester to the second semester, attrition brought the total working number of students

for E down to 53 and for C down to 40.

Before the experiment began, χ^2 frequency analyses were run on certain "face-sheet" information: sex by methods ($.50 < P_{2\text{-tail}} < .70$), class (freshman, sophomore, junior, or senior) by methods ($.50 < P_{2\text{-tail}} << .70$), and curriculum (liberal arts, business, or engineering) by methods ($.30 < P_{2\text{-tail}} < .50$). Thus, the two sets of classes did not differ significantly on various categorizing criteria. For immediate achievement from the first semester, Laidlaw's covariate was a 120-item objective test made by him on both verbal ability and knowledge of psychology (corrected split-half reliability of .94). The final examination for the first semester was a 150-item objective test made up by Laidlaw (corrected split-half reliability of .96).

For the immediate achievement analyses of the first semester, the investigator states,

> The data for the three course sections in each treatment group were treated as one sample. The combination of data was justified by the determination that the variances on the covariants for sections within treatment groups were homogeneous, and because there was no interaction between instructors and conditions of testing. The variances of the treatment groups on the covariants were homogeneous. The simple analysis of covariance technique was used to test each hypothesis. (Laidlaw, 1963, p.26).

On the first-semester final examination, methods were insignificant ($F < .5$).

In connection with the analysis of immediate achievement from

the first semester, one might ask why the investigator went ahead with the one-way analyses after he apparently went through all the labor of the more informative two-way (methods by classes--or, what is the same, instructors). This procedure appears to be sophisticated "data snooping." On the other hand, perhaps such a procedure is desirable in that pooling across instructors (or classes) after one knows such variation is negligible, results in a more refined error term when one goes to a one-way analysis with the student as the unit of analysis.

Two 60-item postsemester tests were developed: 30 items came from the first semester final examination and 30 items from the second semester. The first delayed posttest had a corrected split-half reliability of .86, and the second, .84. Five weeks after the end of the first semester, the first delayed posttest was given. After another five weeks, the second delayed posttest was given. Strangely enough, Laidlaw used the first semester's immediate posttest as the covariate for both delayed posttest analyses. On the first delayed posttest, methods were insignificant ($.10 < P < .25$). On the second delayed posttest, methods came out similarly ($.10 < P < .25$).

The investigator concludes, "The study demonstrated that the belief among college teachers and those who write on educational methodology that frequent testing is a useful means for controlling student learning behavior is not well founded." [underlining inserted by reviewer] (Laidlaw, 1963, p. 46).

The reviewer is somewhat skeptical of the procedure of using the first semester's posttest as the covariate for postsemester retention. This procedure appears to be highly suspect, because such a covariate is not independent of the first semester's treatment effects. In fact, such a covariance procedure makes it harder to obtain defferences on delayed retention, since, in effect, one is obliterating the very differences that he is interested in to start with. The covariates in both analyses (immediate and delayed) should have been the same: the very first pretest, which is unaffected by treatment differences. It is true that one wants the covariate to be correlated with the criterion measure, but one also wants the covariate to be independent of the very treatment effects he is trying to measure.

The above study by Laidlaw concludes the first review topic: frequency of testing. Apparently, only one study was ever done in this area at the elementary school level. One study touched upon junior high school, while six studies concentrated on the senior high school. From 1929 to 1963, nineteen studies were conducted at the post-high school level. The results have been inconclusive. Poor experimental design and inadequate statistics have produced largely meaningless results. The second major topic of this review to follow (test grades as an incentive to further achievement) is even less well researched.

REVIEW TOPIC TWO: INFORMAL ACHIEVEMENT TEST GRADES IN RELATION TO TESTING AS A LEARNING DEVICE

The second topic in this review concerns the use of informal achievement test grades as extrinsic motivation; this second topic concerns the idea of the students knowing a test grade will be received that might affect their future success, might push them on to greater heights in achievement. This topic is closely related to the student's concept of "test": how meaningful it is as an incentive. As already stated in the introduction, this second review topic will deal only with test grades as an experimental variable and not with other types of grades and extrinsic motivation (for example, report card term grades, gold stars, awards, citations, and so on.). Since this second review topic is of major concern to the reviewer's dissertation experiment, references will be examined and criticized in detail.

Nonresearch References: Arguing against the use of test grades as incentives, Odell (1928, p. 17) says, "If a teacher is skillful enough to motivate the work of her pupils to a sufficient degree by other means than checking up on their work and appealing to their desire for high marks, less testing will be needed than if it must be employed for that purpose." [underlining inserted by reviewer]

Another negative criticism of test grades as incentives is given by Kneeland and Bernard (1953, p. 499): "But take away the

grading aspect, the positiveness of rights and wrongs--and what is onerous with a grade is suddenly challenging--a game, a matching of wits, a sourcespring of earnest discussion, in which the objective is not defense but learning the truth." The authors also assume that the word "test", even though test grades might not be given, is meaningful enough to students as an incentive in itself. (p.500) : "Objective tests, when not graded, do much to arouse interest and to give variety." This is an interesting problem that has yet to be solved.

Experimental Studies : In contrast with the first review topic, frequency of testing, no breakdown will be made on the basis of educational level because of the dearth of material.

Panlasigui and Knight (1930) found that students given arithmetic drill material in fourth grade without external motivation did more poorly than those given both the drill material and an external incentive. The external motivation used was a progress chart on the wall based on the students' drill-test grades. A total of 56 intact classes from 10 city school systems were used; the cities were located in the West and Midwest in various states. From each school, at least two E groups and at least two C groups were formed.

Each E and C class received the same weekly drill material: 15 problems of mixed type (as compared to "isolated" drill material of only one type). Only whole numbers were used in these previously learned arithmetic skills. The E and C classes received the drill

materials only one day a week for 20 weeks. The E treatment consisted of the use of individual progress charts and class progress charts. A pretest and posttest were given to all 56 classes.

Out of a total for all E and C classes of 988 students left at the end of the experiment, 358 matched pairs were possible for analysis; the matching was on the basis of pretest scores. On both E and C analytical groupings, six was evenly divided. Simple z statistics were used throughout the analysis. On the posttest, $E > C$ ($P_{2\text{-tail}} < .001$). The investigators claim (p. 614): "A just interpretation of the omitted data warrants the statement that the gains were a bit slow in appearing, but that with increasing sureness the Experimental Group responded more successfully as time went on. In other words, the novelty of the progress chart idea did not stimulate a spurt of effort which then tended to die out, but, rather, an opposite effect appeared."

Breaking the analysis down finer, on sex it was found that $E_B > E_G$ (not significant but no definite probability value was given). Then, dividing the 358 pairs into four quarters of ability on the basis of pretest performance, for the top quarter, $E > C$ ($P_{2\text{-tail}} < .001$). For the lowest quarter, $C > E$ ($P_{2\text{-tail}} = 0.86$). Although the investigators did not give the exact results for the middle two quarters, it was said that the second highest quarter was still significantly in favor of E. Panlasigui and Knight (1930, p. 615) conclude, "The beneficial effect of awareness of success, then, was substantially in direct proportion to the amount

of success available for motivation."

Fay (1937) studied the effects of two test grading systems upon undergraduate juniors and sophomores in introductory psychology. At the start of the experiment, 196 subjects were available. The large class was broken up into an E group of 89 and a C group of 96. E used the "open" test marking system (to be distinguished from the less related semester marking system); each student could find out his monthly test grade and final examination grade in terms of A, B, C, D, and F. The C group used the "closed" test marking system; after each test, they could find out their test grade only in terms of P^+ (satisfactory), D, and F.

Both groups were taught by the lecture and quiz method. Two classes a week were devoted to lectures; the lectures were handled equally by two instructors. In a somewhat contrived manner, once every week the two groups were broken down into eight quiz and discussion sections. Once every fourth week, a 125- or 150-item objective test was given. The immediate posttest consisted of a 400-item objective test.

For purposes of analysis, the groups were matched on both percentile rank on the American Council on Education Psychological Examination and score on the first monthly test in the course. Six separate analyses were carried out: four monthly tests, the final examination and the difference between the first monthly test and the final examination. Within each of the six analyses, ability

categories of A, B, C^+, and C were formed on the basis of the first test given in the course. Unfortunately, no information at all was given about the categorizing test; one hopes that it was different from the four monthly tests used in the analyses, since one needs a control variable to be independent of the criterion measures. Throughout the six analyses, simple z statistics are used; separate two-way analyses of variance would have been more appropriate. The sixth analysis (difference between first monthly test and final examination) will not be considered in the presentation of results in this review, since the calculations appear to be invalid. Between the two tests, one has reflections in his difference scores of different difficulty levels and different content materials. This criticism is especially applicable to standard raw scores (Hull's formula: $X = K + SX_1$) used in all analyses; even if normalized scores had been used, the reviewer would still doubt the validity of the calculations.

Because of the university's academic attrition policy, after eight weeks the E and C groups consisted only of students having an average better than D. Hence, all analyses throughout the whole experiment were conducted only on such subjects. For students of A ability, E > C for each of the four monthly tests; only the third test approached significance ($P_{2\text{-tail}} = .11$). On the final examination, students of A ability yielded E > C ($P_{2\text{-tail}} = .002$). For students of B ability, on all four monthly tests, C > E; only the second, third, and fourth monthly tests are worth discussing:

$P_{2\text{-tail}} = .04$, $P_{2\text{-tail}} = .16$, and $P_{2\text{-tail}} = .02$, respectively. On the final examination, students of B ability yielded C > E ($P_{2\text{-tail}} = .008$). For students of C^+ ability, contradictory results were obtained: E > C on the first and third monthly tests, while C > E on the second and fourth monthly tests (all results highly insignificant). On the final examination, students of C^+ ability yielded C > E (again, insignificant). For students of C ability, C > E on all four monthly tests. Only the first and fourth monthly tests are worth mentioning: $P_{2\text{-tail}} = .16$ and $P_{2\text{-tail}} = .02$, respectively. On the final examination, students of C ability yielded E > C (not significant).

Fay (1937, p. 551) rationalizes his rather contradictory results: "In other words, if students securing an A on the first test knew their marks, they apparently put forth extra effort to retain their positions. The C students in the experimental group attempted to improve their standing. The B and C^+ students, on the other hand, were apparently satisfied, did not unduly exert themselves, and consequently declined relative to the rest of the class."

Bostrum, Vlandis, and Rosenbaum (1961) claim that experimental evidence in realistic classroom situations is rare on the problem of providing extrinsic motivation (test grades, gold stars, prizes, and so on). The investigators studied changes in attitudes when reinforced by randomly assigned test grades. A total of 228 undergraduate students in communication skills classes were the subjects. All subjects were given an attitude questionnaire consisting of four

10-item scales (federal aid to education, legalized gambling, capital punishment, and socialized medicine). The scales on federal aid to education and capital punishment were not used in the analyses because of nonsatisfaction of measurement assumptions. About six weeks after the attitude questionnaire was given, all subjects wrote a half-hour essay on either legalized gambling or socialized medicine. Each subject was assigned the topic he had evidenced the strongest attitude on the questionnaire six weeks earlier. Marks of A, D, or No Grade were randomly assigned to the essays. The next class period the essays with the randomly assigned grades were returned to the subjects; during that period, each subject was given both attitude scales (legalized gambling and socialized medicine). Bostrum et al. (1961, p. 113) say, "Finally, subjects were asked to indicate their satisfaction with the essays."

The final analyzable sample consisted of 127 students. The analysis for mean attitude change (from pretest questionnaire to posttest questionnaire) yielded heterogeneity of variance; hence, the investigators used Cochran's approximate t method for testing the significance of the difference between means of change scores: $A > D$ ($P < .01$); $A >$ No ($P < .05$); and $D >$ No ($P > .10$). This analysis was only for the one posttest attitude scale (out of two different posttest scales) related to the essay written by that particular student. It should be noted that Cochran's procedure was inappropriate here, since the investigators had three samples, not two; therefore each of the three comparisons fails

to take into account overlapping variance. An analysis of variance should have been used in connection with a stabilizing transformation of scores.

Continuing with the analysis, Bostrum et al. (1961, pp. 113-114) say, "An analysis of mean change in relation to initial position indicates that those who had initially assumed a favorable position on each of the issues . . . changed significantly more (. . . $p < .01$) than those who were unfavorable." Further, Bostrum et al. (1961, p. 114) claim, "By comparing the change scores of subjects who had written an essay on a particular topic with those who had not written on that topic . . . [it was found that] This difference is significant (. . . $p < .01$) suggesting that the writing of an essay, independent of grade received, produced change in attitude."

Satisfaction with grade received was also investigated. A χ^2 analysis was performed on the distribution of responses (satisfied and not satisfied) for each of the essay grade levels (A, D, and No Grade). It was found that $P < .001$. Finally, it was concluded (p. 114), "The results suggest support for the hypothesis that a 'good' grade serves to reinforce the behavior for which it has been administered."

Hawk and DeRidder (1963) also claim that actual experimental evidence is rare regarding incentives in realistic learning situations. The investigators used two grading procedures with college students. In one section, students' course grades were determined in the usual way: by course test performance and a term project.

In the other section, the students' course grades were determined by their cumulative grade-point average computed up to the start of the course.

Four sections of educational psychology at the university level comprised a total of 118 subjects; most were sophomores and juniors. Each of two instructors taught two sections. Hawk and DeRidder (1963, p. 548) say, "Each section met three times each week for a 50-minute period, Monday, Wednesday and Friday." The assignment of teachers to sections had already been determined by the administration, probably in a nonrandom manner. Also, it appears that a nonrandom procedure was used to select the E section from the two sections of each instructor. To combat the contaminating effect that instructors' knowledge of differences in methods might have on the effectiveness of teaching, the students were told of the experiment but a deliberate attempt was made to keep the instructors ignorant of the procedures. It seems to the reviewer that the subjects should also have been kept ignorant of the fact that an experiment was under way. Two 60-item, multiple-choice unit tests and a 100-item, multiple-choice final examination formed the criteria.

Hawk and DeRidder (1963, p. 550) say, "Mean scores for the two groups taught by each instructor were tested for possible significant differences, but all differences between instructors were so slight as to be insignificant." However, no details are given as to exact results or to methods of calculation and how they tie in with later statistics. Then, using t tests with 116 df, on the

first unit test, with E representing the predetermined grade group, $C > E$ $(.01 < P_{2\text{-tail}} < .02)$. On the second unit test, $C > E$ $(.01 < P_{2\text{-tail}} < .02)$. On the final examination, $C > E$ $(P_{2\text{-tail}} << .001)$. On the case study, $(.05 < P_{2\text{-tail}} < .10)$. Hawk and DeRidder (1963, p. 550) conclude, "Such findings raise questions about the validity of arguments of many educators that grades destroy motivation."

Nolan (1964) performed an experiment very similar to that of Hawk and DeRidder (1963). Nolan also studied the arbitrary assignment of grades to an essay test but this time with respect to the effect on subsequent test performance rather than attitudes. Two intact classes in undergraduate educational psychology were used. The class with 99 students met Monday, Wednesday, and Friday at 9:00 A. M., while the class with 126 students met the same days at 10:00 A. M. In the total body of students, there were 68 males and 157 females. The course carried six credits' weight rather than the usual three. Each class day, one hour was spent in a large lecture class where two instructors taught as a team; the second hour was spent in small discussion sections where graduate assistants were in control.

During the third week of the semester, on a Monday, it was announced that an essay test would be given that Wednesday. A three-question essay test was then given on Wednesday. The papers were graded outside of class by the instructors and returned at the next meeting (Friday). Nolan (1964, p. 36) states, "Those students for

whom there was no grade-point-average, were assigned grades of B and D equally so that it would appear to the students that the full range of grades had been used." No discussion of results at Friday's class was permitted, and the papers were returned to the instructors at the end of class. On the following Monday, an 18-item, 5-choice criterion test was given; it was found that the KR 21 coefficient was .81. This objective test covered the next assignment and was not directly related to the content of the essay test.

One control variable was cumulative grade-point average (GPA). A second control variable was to have been the score on the Work Persistence Attitude Scale (WPAS). The 20-item, 5-point-continuum type WPAS was given at the first class meeting. Ten items dealt directly with work persistence attitudes, while the other ten items were distractors chosen from a standard personality test. However, it was found that the WPAS possessed almost no discrimination power, and the internal consistency coefficient was only .32 (computed by analysis of variance). Hence, WPAS was discarded as a second control variable.

Three separate analyses of variance were performed: all subjects as a whole, females only, and males only. It seems to the reviewer that, if separate analyses by sex can be performed, then it would have been better to include sex as a second control variable in the whole-group analysis; then measures of interactions with sex could have been gotten. For all subjects, as one might expect, the control variable of GPA was highly significant ($P < .005$); all other

main effects and interactions were highly insignificant ($P > .25$). For females alone, GPA differed significantly ($.005 < P < .01$), while treatments and treatments by GPA were insignificant ($P > .25$). Strangely enough, for males alone, the control factor of GPA was highly insignificant ($P > .25$), while treatments were less insignificant ($.10 < P < .25$)! Again, however, treatments by GPA were insignificant ($P > .25$).

The investigator suggests several limitations of his study: the failure of WPAS to function as a control variable, the inability to administer treatments more than once because of ethical considerations, and the possibility that college sophomores are so ingrained with school procedures that a single essay quiz does not make much difference one way or the other. Nolan (1964, p. 61) lists possibilities for further research: "The development of a sensitive and sophisticated instrument for the measurement of student attitudes toward the grades they receive is needed. . . . Studies are needed at various age and experience levels in order to determine where the grades assigned to students' work are most influential." The latter recommendation has strong relevance for the reviewer's dissertation experiment.

REVIEW TOPIC THREE: TEST CORRECTION WITH RESPECT TO INFORMAL ACHIEVEMENT TESTING AS A LEARNING DEVICE

The detailed rationale for treating test correction method as an aspect of informal achievement testing as a learning device has already been presented in the general introduction. No more need be said here. Test correction will be reviewed only very briefly, since it is not of concern to the reviewer for his dissertation. The only other topic in this review that will be done in detail is the tenth one : student attitudes toward informal achievement tests. Each of the briefer review topics (of which "test correction" is one) will still be separated into nonresearch references and experimental studies; in turn, experimental studies will be broken down only by broad results: positive, negative, or no difference.

Nonresearch References: Potential methods of having students correct tests have been described by various writers: Jeep (1933), Smeltzer (1933a), and Lee and Symonds (1934). With respect to objective tests, Davis (1943, p. 530) says, "The burden of correcting short tests and written exercises may be shared by pupils in scoring the papers." Krause (1966) presents a similar argument.

Experimental Studies: Gates (1921) has found favorable results for the student-corrected mode. Cocks (1929) found that tenth-, eleventh-, and twelfth-grade boys who corrected their own test papers in physics did much better than the groups in which the

usual teacher-correction procedure was used; girls were not used in the experiment. Cocks also found that, among the members of the student-correction groups, the younger, less intelligent ones benefited most. He obtained similar results in content areas other than physics. Buckner (1931) found a slight difference in favor of the student-corrected-test group over the traditional teacher-corrected-test group in a high school foreign language course. Curtis and Woods (1929) found the student-corrected mode was superior to the teacher-corrected mode in seventh-grade general science, eighth-grade general science, ninth-grade biology, tenth-grade biology, and eleventh-grade chemistry. Curtis and Darling (1932) replicated the 1929 experiment and found the same results. Finally, Curtis (1944) found similar results in high school science; unfortunately his results were confounded with feedback mode.

REVIEW TOPIC FOUR: TEST RESULT FEEDBACK AS RELATED TO TESTING AS A LEARNING DEVICE

This review is confined to the use of traditional teacher-made tests used in realistic classroom situations; the reviewer will omit the voluminous literature from contrived laboratory studies on immediate and delayed reinforcement in trivial tasks with mechanical apparatuses. Further, the investigations on programed instruction dealing with feedback response frame schedules will not be considered here. Such peripheral topics have been adequately reviewed elsewhere. Finally, it might be said that feedback and correction are very closely related and perhaps might be treated more appropriately as a single topic; however, for purposes of efficiency and analysis, the two topics have been kept separate in this review.

Nonresearch References: Opdyke (1927, p. 36) says, "If it [that is, the frequent teacher-made test] can be kept short enough to permit children to finish and then to discuss it with the teacher in the same period it will have an immediacy of impression and effect that will prove invaluable." Symonds (1927, p. 533) claims, "One advantage of the new-type test is that it may be immediately scored and discussed, thereby making the most of the discussion when interest in the test is running high. . . ." Weber (1929, p. 537) maintains, "Whether given at the end of the term or at the close of a certain unit of work, their results should be reviewed with the entire group. . . ."

Kneeland and Bernard (1953, p. 499) claim that teacher-made,

objective tests are often overlooked as to their value "To stimulate more and better discussion." Koester (1957) described a loose, non-control-group tryout "experiment." He made use of small discussion groups for talking about test results as compared to the traditional intact-class, posttest discussion sessions led by the teacher; he claims that the latter is not as effective as the former. In relation to a loose, noncontrol-group tryout of true-false tests, Flook (1959, p. 262) claimed, "In short, use of the test had made a valuable contribution to the course, in particular by improving the quality of discussion [afterwards]."

Tyler (1959, p. 15) said, "Another policy which can increase the positive values of testing is to use similar tests periodically throughout the instructional program and to review with the students their performance on each test. . . . This practice also reduces the emotional tension surrounding testing. Testing becomes a natural part of the total learning process rather than an infrequent and traumatic experience." Anderson (1960, p. 51), in discussing the use to which classroom tests are often put, said, "Reinforcement. . . has come in for little specific attention. . . . the periods of time which elapse between the student's response and some of the meager reinforcements he does receive are frequently so long that most of the effect is destroyed." Coladarci (1964, p. 258) claimed, "Testing, as a part of the evaluation of the behaviour [sic] in relation to the goals, must parallel the educative process in order to provide feedback on the progress being made."

Finally, Dyer (1967) provided a learning theory model for instructional feedback in terms of instructor, student, student environment, feedback, and points in time.

Experimental Studies: One of the earliest realistic feedback learning experiments was that of Book and Norvell (1922). They found that feedback in the form of immediate numerical scores (as opposed to letter grades) in simple arithmetic tasks was superior to no such feedback (simple practice without knowledge of results). Brown (1932) studied feedback in fifth-grade and seventh-grade children in a large city school system in connection with previously learned arithmetic skills. In each grade, one group employed immediate feedback in the form of a bar graph, while the other group got delayed feedback after teacher correction. In both grades the bar-graph method was most effective. Ross (1933) studied feedback from the standpoint of how much knowledge of results is given each student. Ross found no difference among four degrees of detailedness in a tests and measurements class at the college level.

Plowman and Stroud (1942) found that subjects who got test result feedback in the form of written teacher solutions of wrong problems reduced their errors by half a week later. Krueger (1947) performed an experiment comparing students' honesty in reporting grading errors under different conditions. Angell (1949) dealt with immediate knowledge of results in college chemistry at the freshman level by means of a punchboard. The punchboard group was superior to the usual machine-scored, delayed feedback group. Jones

and Sawyer (1949) again found superior results for immediate feedback by means of the punchboard as compared to traditional delayed feedback in an undergraduate freshman course.

In Armed Services classroom training, both Stone (1955) and Bryan and Rigney (1956) found that the more complete the amount of feedback (for example, discussion of errors versus simple return of numerical scores), the better the achievement. Page (1958b, p. 173) claims, "Each year teachers spend millions of hours marking and writing comments upon papers being returned to students, apparently in the belief that their words will produce some result in student performance, superior to that obtained without such words. Yet on this point solid experimental evidence, obtained under genuine classroom conditions, has been conspicuously absent." He performed a tightly controlled experiment with 74 intact classrooms in seventh through twelfth grades in several content areas. Three treatments were used: free comments, specified comments, and no comments. A monotonically decreasing performance level was found for the latter order of treatments.

Sturges and Crawford (1963) studied immediate versus delayed feedback with realistic, factual material. Sassenrath and Garverick (1965) found that the teacher-led discussion method of feedback was superior to looking up wrong answers in the textbook, checking over answers from correct ones on the board, and no feedback. The subjects were students in undergraduate introductory psychology. Paige studied feedback in eighth-grade mathematics students. E received

immediate feedback in the form of special carbon copies of their tests which they kept after turning in their original tests; C had to wait as usual until the next day. E exceeded C in performance.

Daniel (1968) studied feedback in undergraduate educational psychology classes. E received immediate knowledge of results on a teacher-made test; E had to look up correct answers to their mistakes. C received knowledge of results a day later. Both E and C received discussion of results. Strangely enough, the delayed feedback group excelled over the immediate feedback group. Daniel and Witchel (1967) found similar results for one-week delayed feedback over immediate feedback in college students on teacher-made tests.

REVIEW TOPIC FIVE: PRETESTING AS AN ASPECT OF TESTING AS A LEARNING DEVICE

Considered positively, pretesting as an instructional device might be thought to involve benefits to both the teacher and the student in terms of diagnosis and, especially for the student, of structuring in his mind subsequent subject matter. Considered negatively, pretesting as a methodological device can taint subsequent measurements by the phenomenon of sensitization.

Nonresearch References: Breslick (1921, p. 277) says, "The old type [of examination, namely, the essay test] is worth more in diagnosing pupil difficulties. The extent of the value of the new type [that is, the objective test] in diagnosis is yet to be fully demonstrated in history." Spencer (1923) talks about informal achievement tests used as diagnostic devices in high school algebra. Horn and Ashbaugh (1926) recommend the use of the test-study method in teaching elementary school spelling. Cody (1929), Weber (1929), McGinnis (1929), Jones (1929), and Burr (1929) discuss the diagnostic values of testing. Breed (1930) recommends the use of pretesting in teaching elementary school spelling.

Horn (1933) is also in favor of the test-study method of teaching elementary school spelling. Hutchinson (1933, p. 436) said that tests ". . . should help the student organize his knowledge . . . [and] should give aid to both student and teacher in diagnosing the weaknesses . . . in the student's knowledge." He was one of the

few writers to see the structuring benefits of tests. In relation to diagnosis, Smeltzer (1933a, p. 527) said, "Much classroom testing is of little value from a teaching or learning standpoint because no further analysis is made of the test results" Davis (1943, p. 528) claimed, "A desirable use of a test is to survey at the beginning of a subject the pupil's previous background and the extent to which any abilities have already been developed." Lockhart (1948) suggests using pretests to find out how much students already know. Kneeland and Bernard (1953) said much the same thing.

Experimental Studies: Kingsley (1923) found that the pretest method was superior to the study-test method of teaching spelling. Kilzer (1926) also studied the pretest method of teaching spelling versus the study-test method. The pretest method was superior. Watts (1928) was another to study the test-study method of teaching spelling. Jersild (1929) performed three experiments on pretesting versus no pretesting. With multiple-choice tests and essay tests, positive results were obtained for pretesting, while pretesting with true-false tests gave negative results. Kirkpatrick studied the effect of pretesting in high school physics. The pretested groups were superior to the nonpretested groups. Keys (1934b), dealing with upper classmen and graduate students in educational psychology, found that for items on which subjects were pretested at the start of a unit of instruction, rates of achievement were higher than for nonpretested items.

Gates (1939) studied the pretest method of teaching spelling with the usual techniques. In spelling, for third through eighth grades, the test-study method was superior to the study-test method. Luce (1939) studied seventh, eighth, and ninth grades on a geography passage on the basis of the methods of test-study, study-test, and just study. The study-test group appeared to do slightly better on all posttests given to the three groups. Tiedeman (1948) found that subjects in fifth-grade pretested on geography passages were slightly superior on achievement than *non-pretested groups*.

Finally, several investigators have considered the methodological aspects of pretesting: Solomon (1949), Hovland, Janis and Kelley (1953), Piers (1955), Lazarsfeld (1957), Campbell (1957), Lana (1959a, 1959b), Lana and King (1960), Entivisle (1961), Campbell (1963), Edling (1963), and Rayder and Neidt (1964).

REVIEW TOPIC SIX: RETESTING AS AN ASPECT OF TESTING AS A LEARNING DEVICE

This section is directed toward those studies that have made attempts to compare performance of subjects who are retested several times with informal achievement tests, with those who are not, on the criterion of delayed achievement. It should be noted that many studies have been done on recall and retention, but most have dealt with trivial materials in unnatural, laboratory-like situations; this review covers only the realistic learning situation experiments that approximate the classroom setup.

Nonresearch References: Spencer (1940, p. 14) provides an excellent description of the matter at hand:

> There are exceptions to the assumption that all retention curves show a drop after the learning performance is ended. Ballard [1913], by an experiment in which pupils memorized a poem, found by retests day after day, that the scores went up during a period of five days following immediate testing. Ballard designated this process, which is opposite to forgetting, as reminiscence. It follows that for there to be an increase in the amount of retention at delayed recall, the material must have been incompletely learned originally. If recall at the close of learning has been complete no later recall can be greater; hence there can be no reminiscence.
>
> Woodworth [1938] terms the idea that a forgetting curve can rise is [sic] absurd and that it is impossible for one to retain more than was learned originally unless some other process enters to produce added learning. Woodworth advances the possibility that reminiscence is due to the involuntary or voluntary reviewing of the material by the subjects. Ballard admits this possibility but does not believe that the whole

> effect of reminiscence could be so explained. [underlining inserted by reviewer]

Experimental Studies: Yoakam (1922) found that testing retards the curve of forgetting more than just rereading written passages on which the tests are based. Jones (1923) was one of the first to study the reminiscence phenomenon to any great depth with realistic learning materials at the college level. In general, he found that retesting (in many different schedules) impeded the curve of forgetting as compared to nonretesting conditions. Spitzer (1938, 1939) found that, for geography passages with sixth-graders, the closer retests were to the initial posttest, the greater the curve of forgetting was retarded. Spencer (1940) replicated Spitzer's experiment, this time presenting the geography orally. Similar results were obtained.

Sones and Stroud (1940) studied retesting versus simple rereading in different time schedules with a passage on geography for seventh graders. When retesting was used relatively close to the first posttest, it was superior, but rereading exceeded retesting as one got further away from the initial posttest. Davis and Rood (1947) found the reminiscence phenomenon in arithmetic for three testings with the same test. Little (1960) studied reminiscence in undergraduate biology students; the phenomenon was again present in the case of retesting versus nonretesting. He also discusses methodological issues with respect to the calculation of reminiscence scores. Finally, Celinski (1968) studied

retesting in graduate level electronics courses in two different universities; unfortunately, no control groups were used. Short, announced, repetitive tests were given quite frequently throughout the course. Each student moved at his own rate, but to progress further, he had to obtain a perfect test score; if he did not, he kept taking the test over until this occurred. Each subject evidenced increasingly better performance as the semester progressed.

REVIEW TOPIC SEVEN: TEST EXPECTATION AS AN ASPECT OF TESTING AS A LEARNING DEVICE

The motivational aspects (and hence, the additional learning benefits) of test expectation (perhaps in the form of prior announcement of an impending test) have already been discussed in detail in the introduction. This review section gives a brief overview of studies in this field.

Nonresearch References: Gable (1936, p. 1), in connection with test expectation, said, ". . . educators seem to be agreed that pupils tend to accomplish more when confronted with the realization that a day of reckoning is at hand when they are expected to give an account of their knowledge. Such a situation contains dynamic or motivating properties which aid the crucial aim in teaching--motivation."

Experimental Studies: Jones (1923), as part of his extensive series of retesting experiments at the college level mentioned above, also did significant work in this area. At the beginning of the hour-long lecture period, half of each class used was given a slip of paper that notified them of a five-minute quiz at the end of the period on the material of that day's lecture; the other half of each class received a "dummy" library notice on a similar slip of paper. The results for pooled unexpected groups were almost identical. Schutte (1925) used "normal school" students in an introductory education course to measure the effect of announcement

of an impending test. The experiment was conducted twice (once for each of two academic years). Two intact classes a year were used: one class expected a final examination, while the other did not. The results for the two separate years were pooled with "methods" still being distinguishable. The expected group did superior to the unannounced group.

Pease (1927, 1930) studied the effects of "cramming" and expectation versus nonannouncement. Both high school and college subjects were used. On the day the test was to be given, E was told of the impending test and was instructed to "cram" for it in the time set aside for this purpose; C was given the test immediately without announcement or time for equivalent "cramming". E did much better than C on both immediately and delayed retention. White (1932) compared a group that was told that a final examination would be given in the course and what types of material would be on it, with a group that was not told to expect a final examination. The expected group was superior on the final test's performance.

Corey (1935) was one of the first to do work in this field with realistic learning materials and environment. Gable (1936, p. 5) said Corey ". . . compared correlations of Army Alpha scores of 104 students with test results obtained from surprise quizzes, on the one hand, and on the other hand with results obtained from a final examination announced long in advance. His assumption is that achievement is motivated much more adequately in a final examination

than in a surprise quiz." Gable (1936) studied subjects in ninth-grade biology. She compared three oups: a "pop"-quiz one, a preannounced one, and a nontest control one. She concluded that a mixture of announced and unannounced quizzes is the most effective procedure.

REVIEW TOPIC EIGHT: TEST EXEMPTION AS AN ASPECT OF TESTING AS A LEARNING DEVICE

The reader will recall from the introduction that test exemption has potential motivating properties in at least two ways: exemption from course work by superior test performance, or exemption from tests themselves by superior course work.

Nonresearch References: Odell (1928, pp. 51-52) seriously questions the motivating properties of the process of exemption from the final test by classroom performance: ". . . it comes to be looked upon by pupils as more or less of a disgrace to have to take examinations. . . . The whole exemption system tends too much to make the examination a penalty and a disciplinary device rather than an integral and educative part of the instructional process." Cole (1929, p. 120) said:

> It was a rule in the Seattle schools for some years to excuse all pupils of advanced standing from taking examinations. In other words, we made examinations a penalty, and it was considered somewhat of a disgrace to take them. As one teacher put it, 'The only pupils in this school who are taking examinations are those who will profit the least by taking them.'

Nickerson (1929, p. 253) asked:

> Should exemptions by made? I grew up in a system where it was the rule to be excused from examinations if you made a certain monthly grade. Everyone strove to attain that average, and made much better grades than they would otherwise have done. Oh, what a joy it was to get a few days' vacation! I think the teachers really enjoyed not having those extra papers too. Examination under that system became a penalty rather than a

> 'learning exercise.' All would have been well had I not gone to college and had to take examinations. I had never had to take examinations, and just the thought of them appalled me, and still does.

Webb (1929, p. 282) also argues against exemptions from tests:

> The training which students get in correct written expression through tests or examinations is valuable for all classes of students, therefore not any should be exempt from examinations. Graduates of high intelligence are oftimes handicapped in making good in important positions because they failed to acquire the habit of producing elegant and accurate oral and written expression. Some of them were exempted from examinations. [underlining inserted by reviewer]

While the latter part of Webb's passage may be somewhat unscientific, his initial point is well taken, as was Nickerson's point that examinations are an integral part of our lives and to exempt people denies them necessary practice in examination-taking skills. However, on the other hand, the motivating properties of exemption cannot be dismissed lightly.

Finally, in reviewing the state of research in this area in his time, Davis (1943, p. 533) claimed, "Investigations dealing with the effect of exemption from the final comprehensive examination, the extent to which it provides information additional to that the teacher already has by the time it is given, have not yielded answers sufficiently conclusive for generalization."

Experimental Studies: Morley (1926) noted that, while superior students were not affected one way or the other by the exemption

procedure, the mediocre students gained more than they would have otherwise. Engelhart (1931) described an experiment dealing with exemption from the final test by course performance He concluded that this type of exemption appeared to raise the performance of otherwise average students. Gould (1932) performed a survey of 125 secondary schools in 48 states. He reported (p. 145), "No exemptions are permitted in forty-seven of the sixty-one schools requiring final examinations. Pupils are exempted for many different reasons in the remaining fourteen schools."

Meltzer (1933) performed an experiment dealing with exemption from a portion of course work every week on the basis of weekly tests. The exemption procedure was superior to the tratitional nonexemption one. Smeltzer (1931 and 1933b) performed a study in college chemistry very similar to the study of Meltzer (1933). Smeltzer found that the extremes of the ability range were affected very favorably by the weekly exemption procedure in relation to the forced-attendance group. Remmers (1933) made a study of undergraduate engineering students. E was allowed exemption from the final examination on the basis of superior course performance, while C had to take the final examination regardless of previous achievement. E was superior to C on immediate retention but not on delayed retention. Finally, Dole (1951), however, provided evidence on the procedure of exemption from the course by test performance. In his study at the

college level, he concluded that the procedure was very effective.

REVIEW TOPIC NINE: STUDENT PREPARATION FOR TESTS AS AN ASPECT OF TESTING AS A LEARNING DEVICE

A few theorists and investigators have been interested in the ways that subjects prepare outside of class for informal achievement tests. Not much has been done in this area, mainly because it is difficult to control extra-class variables. However, most of the nonresearch references on motivation arising from tests refer directly or implicitly to this external preparation factor. Thus, any vague reference to the simple motivating power of a test (without specifying just how such motivation is brought about) usually implies that the subject has been urged onward outside of class to greater preparation for the test. Such vague test motivation references are considered in this section of the review.

Nonresearch References: Weber (1929, p. 62) says the informal achievement test ". . . teaches the student to express his knowledge accurately and concisely. As a preparation for this expression, the examination, if effective, requires a careful study and review of the course pursued." Pyrtle (1929, p. 119) claimed, "Tests when properly given are a stimulus or challenge to a student to more effective or more thorough work." Weeks (1929, p. 281) asserted that informal achievement tests ". . . act as a motivator. The knowledge of a judgment day seems to keep some folk on the straight and narrow way. Pupils who know beforehand that tney are going to be held accountable for a given unit of subject matter will study more diligently, all other things being equal." Similar ideas have

been expressed by Colvin (1913), Symonds (1927), Pearson (1929), Cole (1929), Verables (1929), and Krause (1966).

In connection with the general problem of motivation from tests, Ruch (1929, p. 10) said, "It is unfortunate that we have so little direct information as to the motivating effect of examinations. That examinations do have this value has been tacitly agreed but never proved." Finally, Tyler (1959, p. 10) claimed, "Well-motivated students have commonly put extra time and effort into study when they thought they were soon to be tested."

Experimental Studies: Douglass and Tallmage (1934) made an intensive effort to discover how subjects prepare for examinations; so also did Meyer (1934, 1935). In all cases it was found that the announcement of an objective test produced different study methods than the announcement of an essay test. Class (1935) also performed one of the few major studies in the area of type of preparation used for tests. Briefly, he found that most subjects used very different study habits for true-false tests as compared to essay tests, when such tests were announced in advance. Further, he found that subjects performed best on both true-false tests and essay tests as compared to other types (completion, multiple-choice, and so on) when type of test was not announced in advance. Finally, Vallance (1947), studying senior high school students, again found a strong difference in study methods used for essay tests as compared to objective tests.

REVIEW TOPIC TEN: STUDENT ATTITUDES TOWARD INFORMAL ACHIEVEMENT TESTS

This section includes miscellaneous attitudes of students toward informal achievement tests. Although emphasis is usually on measures of performance and achievement in connection with studying testing as a learning device, measures of attitude are also very important. In connection with attitudes, test anxiety will also be considered in this review section.

Deputy (1929), whose study on frequency of testing has already been described above, also investigated student attitudes toward informal achievement tests; no doubt he was one of the first in this respect. He made a survey of attitudes near the end of the semester after the procedures had been changed (the reader is advised to look over the description of Deputy's experiment in the first review topic). The respondents were told to remain anonymous. Deputy gives only rough percentage statistics but no tests of significance: in E_1, 86% preferred daily written work; in E_3, 85%; and in E_2, 46%. Deputy (1929, p. 333) comments: "at three different times soon after Section 1 had been changed from an experimental to a control section, the students as a class asked to continue the daily written class work, their score in the mid-semester examination was not so gratifying as that of Section 1. . . . It is suggestive to know that the extent of the unfavorable attitude of Section 2 toward the written exercises is due to the fact that their written work came during the second half of the semester, after a half

semester of only oral recitation work."

Turney (1931), who also did a frequency of testing experiment described earlier, gave a questionnaire of yes-no type to the experimental group (the one that underwent frequent testing) at the last class meeting. Out of 41 subjects who took the questionnaire, Turney claims most were favorable toward short, frequent tests. He provides no exact data or statistical tests and did not give the questionnaire to the control group.

Kitch (1932), the major part of whose experiment was also described earlier in the frequency of testing section, also studied student attitudes. A questionnaire of yes-no? type was given to the students at the end of the experiment. Although no tests of significance were made on the frequency data, χ^2 analyses would be easy to perform on the fifteen questions, since nonoccurrence was allowed for. However, ranking the possible benefits to be gotten from practice tests in order of highest number of positive responses, calling attention to points not noticed was first, needing to study sooner and hence more often was second, aiding in learning key textbook facts was third, and hinting at what the teacher considered important was fourth. An open-ended question was also put on the questionnaire.

Lee and Symonds (1934, p. 174) provided evidence on two student test attitudes studies:

> Students in science prefer objective tests to the essay type according to the data presented by

> Hurd [1929] and Diamond [1933]. Hurd found that physics students on the college level preferred objective tests, largely because such tests covered more ground. Diamond studied the preferences of high school students finding that they also preferred objective tests. He also found that pupils preferred tests made out by other pupils and tests where graphic records of results were kept.

Keys (1934a), whose frequency of testing study was described in detail earlier, also dealt with student attitudes. At both the start and end of the term, he gave the same 30-item, yes-no attitude questionnaire to both groups of his study. Using simple z tests for testing the difference from the start to the end of the term, Keys found two questions to be particularly interesting: a significant increase in the number of students favoring tests given every second, third, or fourth class meeting ($P_{2\text{-tail}} << .001$) and a significant decrease in the number of students favoring tests given only three or four times a semester ($P_{2\text{-tail}} << .001$).

Noll (1939, p. 356), whose frequency of testing experiment was described above, provided rough percentage statistics but no tests of significance: "These replies indicate. . . that about half said they would have enjoyed it [that is, the course] more if there had been occasional written tests, and more than three-fourths stated that they thought they would have learned more if there had been such tests."

Bender and Davis (1949) gave a questionnaire about teacher-made tests to 1040 subjects in 41 secondary schools (public, private, and

parochial); the sample of schools was a proportionately stratified one. Apparently, only tenth through twelfth grades were used. No tests of significance were provided; only loose percentage statistics and informal trends were given. In summarizing their major results, the investigators said (p. 65):

> A highly competitive situation exists for grades in secondary schools; all students wish to be judged fairly and by uniform standards; and students desire to enter a test with no advantage or disadvantage to themselves in comparison with other members of the class.
>
> Students in general consider any particular order of materials in examinations to be of little consequence although they show a decided preference for questions that stress problem-solving ability. Those who are unprepared for a test show a preference for multiple choice and true-false items. When they are well-prepared, their preference is for the essay and completion items. "Cramming" is considered worthwhile by a majority for essay, completion, and problem types of tests although many consider "cramming" worthwhile for all types of tests.
>
> A majority of students prefer difficult tests with ample advance notice (2 to 3 days) to easier tests without previous notification. They also wish to know <u>what</u> a test will cover and the <u>kind</u> of items that will be used. Almost all students desire that the papers be returned promptly with grades and corrections on them. Most students welcome tests as often as once a week. Almost all students worry more or less about all examinations. A few worry to such an extent that they are unable to do their best work on a test.
>
> It is evident that most students work for grades and that they desire to have all of their papers scored and to have all of them

> count as credit toward their final grade. The returns indicate that a majority have an outlook on tests that is sufficiently well balanced and wholesome to serve as a suitable basis for the functioning of psychological principles required for the effective use of tests as learning instruments.

DeLong (1955) conducted an attitude study at the elementary school level. The survey was very loosely conducted and very narrow in scope; all that can be concluded from it are hints for further research of a "tight" type. All the elementary teachers of three school systems were given a 10-question essay (that is, open-end) survey as to how they felt their students reacted to tests as compared to nontesting situations. Also, the investigator had some of his university students go into the same elementary schools on both test and nontest days to observe the subjects' behavior. Finally, over 200 longitudinal case studies from the university's elementary laboratory school were examined for the effects of test-taking. No firm conclusions could be drawn from the whole "experiment" other than that children act differently under a test situation than they do under a nontest situation. The investigator admits that much more research is needed on the emotionality of test situations as compared to nontest situations. The reader should note that DeLong's study is a comparative approach rather than Sarason's "isolated" method; the latter takes whatever the child admits on the TASC to be his degree of test anxiety; the former is perhaps more meaningful in its relativistic approach.

Mudgett (1956), whose frequency of testing experiment was

described earlier, prepared two questionnaires: one for the students and one for the instructors. All daily, weekly, and monthly test groups were given the questionnaires. However, too many objections were put forth by the students to their questionnaire, and it was consequently omitted from the study. Thus, the attitude analysis consisted only of the instructors' questionnaire. In the monthly test groups, the instructors noted that the quality of questions asked in class and subsequent discussion were poor; the instructors attributed this situation to poor motivation because of the experimental treatment of monthly tests. Similar comments were made by the instructors of the weekly test groups. On the other hand, the instructors considered the daily test groups to exhibit superior discussion quality and better motivation; also, even though many subjects objected to daily quizzes at the start of the study, as time progressed more and more subjects saw the advantages in the daily quiz program. Further, instructors felt that all schedules of testing in general, and the daily testing program in particular, aided them with scheduling and preparing lessons and moving along with relatively uniform progress.

Koester (1957), although not a formal study as such, reported on the reactions of fifty students in two graduate-level university classes on frequent testing. He claims (p. 207) that the students' opinions have been ". . . very favorable in terms of interest, motivation, and a feeling of having clarified basic principles."

Selakovich (1962, p. 180) reported: "The students in the

experimental section were asked their opinion on the 'pop quizzes' and there was a near unanimity of opinion favorable to the technique. . . . Most of the students who took 'pop quizzes' felt it helped them learn the basic information required in the course even though the results of the experiment indicate this was not true."

Gaier (1962, p. 561) claims, "A consideration of the methods and techniques employed by students in their approach and preparation for a test situation has not, as yet, brought into being a body of systematic research. In spite of the perfection of testing tools, test situations are frequently perceived by both students and teachers as forms of punishment--mild or otherwise, depending on the difficulty of the testing instrument--rather than a learning experience." Seven intact classes of educational psychology at the university level were used. The subjects were instructed on the response form (p. 561), "Assume that you will receive a letter grade of ["A" or "D"] on the test you are to take. List the specific activities, either on your part or on the part of the instructor, that you feel were influential or responsible in making this grade." This response form was given out on the same day's class as the first quiz of the semester was to be given; the subjects had to complete and return the forms before they were given the actual quiz. The assumed grades of "A" or "D" were distributed as follows: 90 women and 46 men received the "A" forms, while 96 women and 44 men received the "D" forms.

All responses were categorized according to contents in phrases

or sentences. For the "A"-grade response-form subjects, four categories were used to classify responses: (a) "success due to self and self-activities", (b) "success due to teacher", (c) "success due to external factors", and (d) "denial of the possibility of receiving an "A". For the "D"-grade response-form subjects, four similar categories were used to classify responses: (a) "failure due to self", (b) "failure due to teacher", (c) "failure due to internal factors", and (d) "not classifiable". In the reviewer's opinion, this dual classification scheme is the only useful result of Gaier's study; from such a scheme an objective attitude questionnaire could be developed that would tap those attitudes about testing that students think of most.

The reviewer does not put much faith in the validity of Gaier's percentage statistics with respect to the classification of the students' responses according to the above dual scheme. One difficulty in this interesting study is that each student could make as many or as few responses in as many or as few categories as he desired. Thus, when Gaier goes on to compute what percent of <u>total</u> responses were attributed to one reason category, the relativity among individuals (the truly important thing--not the highly variable relativity among individual responses) is destroyed: everything is distorted, because each individual may have over-emphasized one possible category in relation to another. The whole situation is analogous to a very unstructured interview.

Hawk and DeRidder (1963), whose experiment in test grades was

detailed earlier, also studied student attitudes toward informal tests. At the end of the experiment, a faculty member who had not taken part in the conduct of the study went around to all sections and sampled student attitudes. It was found that students in the pregraded groups worked less hard and were much less motivated in general than were those students in the groups under the usual grading procedure.

Curo (1963), whose frequency of testing study was already discussed, dealt with attitudes toward frequent tests by means of a questionnaire and individual interviews. The questionnaire remained anonymous and subjects were deliberately asked to be honest. The 12-item, yes-no questionnaire was given only to the daily test groups. As might be expected, since all twelve questions were clearly directed toward the Hawthorne-producing experimental treatment, most subjects answered in a positive halo-effect sense in favor of daily quizzes. No statistical tests were run; only rough percents were given. No reliability or validity evidence was cited for the questionnaire. Since the interviews were of open-end type, the findings were too divergent to discuss systematically and concisely here.

Test anxiety will also be considered with "attitudes" in this review, since its manifestation is usually measured by a paper and pencil attitude questionnaire. A multitude of studies have been done on test anxiety. However, with respect to the reviewer's dissertation experiment (the learning benefits that result from frequent testing in the elementary school), only Laidlaw (1963) attempted

any measure of test anxiety. He administered the specially developed "Test Behavior Questionnaire" (TBQ) of Hayes (1960); unfortunately, this was a relatively unestablished research instrument whose technical merits are still in doubt. Two equivalent forms are available: A and B. The alternate-form reliability is .63. Each form contains 33 statements of the agree-disagree type. Laidlaw says (p.22), "High scores indicated considerable irrelevant or interfering behavior in a test situation, while low scores indicated little such behavior."

TBQ-A was given at both the start and finish of the frequency of testing experiment (described previously). TBQ-A (pretest) was used as covariate for the criterion of TBQ-A (posttest). Homogeneity of regression was satisfied. However, highly insignificant results were obtained: $\bar{X}_{\text{monthly (adj.)}} > \bar{X}_{\text{weekly (adj.)}}$ $(F << 1)$. Further, the weekly and monthly test groups were pooled and then broken down on paper into high and low ability groups on the basis of the present course's grades. Only the upper 27% and lower 27% were considered. Again, TBQ-A (pretest) was used as covariate for the criterion of TBQ-A (posttest). Homogeneity of regression was satisfied. Marginally significant results were obtained: $\bar{X}_{\text{low (adj.)}} > \bar{X}_{\text{high (adj.)}}$ $(.05 < P < .10)$.

Finally, Laidlaw had asked all subjects in all groups to put down in writing how often they wanted to be tested. By the end of the experiment in the weekly test group, 86% of the students favored frequent tests, compared to 52% at the start. By the end of the experiment in the monthly test group, 66% of the students favored

frequent tests, compared to 64% at the start. No tests of significance are provided.

In connection with test anxiety, Laidlaw tries to rationalize his results (p. 45): "The group that was tested each week was tested four times more frequently than the one tested each month. Each weekly test accounted for a smaller proportion of the total course evaluation, so the risk associated with a weekly test was much smaller. In spite of the difference, the weekly tested group did not learn to cope with tests with less irrelevant behavior under the reduced risk condition."

In connection with the insignificant test anxiety results, the reviewer thinks it also should be noted that TBQ-A was given to both groups by Laidlaw at both the start of the study and at the end of the study. Thus it could have been expected that no significant differences would result, since the students were done with the course, and correspondingly, the fear of tests should have decreased greatly. On the other hand, if the TBQ had been given during the testing process, significant differences (those of "manipulatable process' type, as compared with "predispositional state" type) might have been more readily obtained.

Nolan (1964) also studied attitudes. The failure of his Work Persistence Attitude Scale to function the way a reliable and valid measuring instrument should, has already been discussed above in connection with Nolan's test grading experiment.

REVIEW TOPIC ELEVEN: TEST TYPE AS AN ASPECT OF TESTING AS A LEARNING DEVICE

This topic might well have been combined with "student preparation for tests", the ninth topic. However, there are a few other distinct points the reviewer wants to make in connection with test type other than the inducing of different study habits.

Nonresearch References: Ballard (1925) makes claims that true-false items yield more test learning benefits than do other types of test items. McCall (1920) provides arguments in favor of the objective types of tests over the essay test with respect to didactic value.

Experimental Studies: Remmers and Remmers (1925) compared true-false items with recall (or completion) items; again, according to them, the didactic value of true-false items cannot be denied. Cocks (1929) found superior didactic results for true-false tests as compared to multiple-choice and completion types of tests. However, considering pretesting, Jersild (1929) found true-false items to be didactically inferior to multiple-choice and essay items.

REVIEW TOPIC TWELVE: "TEST-LIKE EVENTS" AS AN ASPECT OF TESTING AS A LEARNING DEVICE

As already stated in the introduction, "test-like events" are included in this review because of their implications for further research in relation to furthering knowledge about testing as a learning device. Rothkopf's "test-like event" procedures provide a tightly controlled environment for investigating more basic issues of tests as to just what it is that causes one to learn more than he would otherwise had he not taken the test. Basically, as already explained in detail in the introduction, "test-like events" are study-guide questions inserted in reading passages or assignments when given in class (that is, this approximates a test situation in its evaluative aspects as compared to study-guide questions given as outside class homework where the study situation is too informal and nontestlike for inclusion here).

Nonresearch References: Cason (1939), acting as theorist rather than experimenter, recommended the use of in-class, study-guide worksheets. Langman (1963, p. 534) offered negative criticism: "Perhaps this passivity in responding to reading materials, expressed by students in requests for syllabi, outlines, and study questions [that is, "test-like events"], is in part the result of our recent teaching methods, which emphasize the provision of external motivation by means of such study materials. Such motivation is artificial."

Experimental Studies: In connection with study-guide questions, Hertzberg, Heilman, and Leuenberger (1932) studied college sophomores in educational psychology. The experiment matched spring semester (E) students with winter semester (C) students. Three different comparisons were made. The difference between E and C was simply that E was given several representative examinations of the course from which they could study throughout the duration of the course. The first comparison was just on the subject matter of the first unit of work of the semester. An examination of just the first unit was given to both E and C; E did significantly better than C. The second major comparison of this study concerned the work of units two through six of the semester. A special examination on these five units was administered to both groups. Again, E did significantly better than C. However, on the third comparison of this study (the final examination), no significant difference was found.

The Motion Picture Research Project (1947, p. 256) dealt with "A procedure which required pupils to participate more actively during the film showing by answering questions [inserted in the film] about various points just after they were presented." Study-guide-question film groups achieved higher results than nonstudy-guide-question film groups. McKeachie and Hiler (1951, p. 224) said, "Every subject matter, be it science, literature, or the arts, is an organized body of knowledge, not a mere array of isolated facts; hence a knowledge of this subject matter should be an

organized structure within the student's mind. Expectations, questions, and problems are intrinsic to all such organized structures." The investigators performed an experiment in elementary psychology at the university level. Worksheets were used in class to guide the independent study of subjects; those who had to complete and turn in the study-guide worksheets were superior to the usual unguided study group on posttests given at the close of the experiment. Robinson (1926), Hurd (1931a,b), Greene (1934), Harrington and Lippert (1934), and Anderson (1942) all performed experiments similar to that of McKeachie and Hiler (1951).

Finally, Rothkopf (1963, 1965, 1966a,b,c, 1968), Rothkopf and Bisbicos (1967), and Bruning (1968) have dealt essentially with completion-type review questions inserted in the text itself; in this respect it begins to approximate programed instruction but is still not the same because of the lack of "framing" and because of the retention of traditional reading passage format. These investigators have left out certain types of words (quantifiers, adjectives, nouns, and so on) and tried to relate this to such things as the degree of relatedness of such omissions to the text or to the real test questions that follow the reading passage. The whole advantage to Rothkopf's procedures is that one can gain a great deal of control in studying an effect such as content structuring in test-like situations, that was hitherto unavailable.

BIBLIOGRAPHY

GENERAL INTRODUCTION

Anderson, S. B. Can tests teach? Cal. J. second. Educ., 1960, 35, 50-55.

Butler, W. F. The value of informal tests in supervision. Yearb. ele. sch. Prin., 1922, 1. Cited by Kirkpatrick (1933).

Elston, B. Improving the teaching of history through the use of tests. Hist. Outlook, 1923, 14, 300-305. Cited by Kirkpatrick (1933).

Fenton, N. New type examinations and their daily use in the classroom. Educ., 1929, 50, 150-158. Cited by Kirkpatrick (1933).

Gardner, E. F. Development and applications of tests of educational achievement in schools and colleges. Rev. educ. Res., 1953, 23, 85-101.

Henricksen, J. Value of objective tests in clothing courses. Pract. home Econ., 1930, 8, 317-334. Cited by Kirkpatrick (1933).

Kimmel, W. C. The use of practice tests in the teaching of the social sciences. Hist. Outlook, 1923, 14, 354-358. Cited by Kirkpatrick (1933).

Koester, G. A. Using instructor-made tests for instructional purposes. Educ. res. Bull., 1957, 36, 207-208.

Lockhart, A. V. Examinations and education. Sch. & Soc., 1928, 27, 725-726. Cited by Kirkpatrick (1933).

McKeachie, W. J. In N. L. Gage (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Obourn, E. S. The improvement of high school physics teaching by a regularly scheduled unit testing program. Sci. Educ., 1932, 6, 497-505. Cited by Kirkpatrick (1933).

Ruch, G. M. The objective or new-type examination. Chicago: Scott, Foresman, 1929. Pp. 10, 25, 145. Cited by Kirkpatrick (1933) and Turney (1931).

Symonds, P. M. Measurement in secondary education. New York: Macmillan, 1927. Pp. 1,533. Cited by Kirkpatrick (1933).

Woody, C. Informal tests as a means for the improvement of instruction. Dept. ele. sch. prin. Bull., 1929, No. 8, 435-442. Cited by Kirkpatrick (1933).

REVIEW TOPIC ONE: FREQUENCY OF TESTING

NONEXPERIMENTAL REFERENCES

Odell, C. W. Traditional examinations and new-type tests. New York: Century, 1928. Pp. 17, 39, 56. Cited by Kirkpatrick (1933).

Opdyke, J. B. Constructive examinations. Educ. Rev., 1927, 73, 33-43. Cited by Kirkpatrick (1933).

Parker, S. C. Methods of teaching in high schools. New York: Ginn, 1920. Pp. 493, 494. Cited by Kitch (1932).

Ragusa, T. Objective test: how it can be used as a method of teaching. Bull. high Pts., 1930, 12, 44-46. Cited by Kirkpatrick (1933).

Ruch, G. M. The objective or new-type examination. Chicago: Scott, Foresman, 1929. Pp. 10, 25, 145. Cited by Kirkpatrick (1933) and Turney (1931).

Wrightstone, J. W. The relation of testing programs to teaching and learning. In W. G. Findley (Ed.), Yearb. nat. Soc. Stud. Educ., 1963, 62, Part II.

EXPERIMENTAL STUDIES: ELEMENTARY SCHOOL

Mann, L., Taylor, R. G., Jr., & Proger, B. B. The learning and test anxiety effects of frequent testing in third-grade arithmetic. King of Prussia, Penna. (443 South Gulph Road): Research and Information Services for Education (Montgomery County Public Schools), 1968.

EXPERIMENTAL STUDIES: JUNIOR-SENIOR HIGH SCHOOL

Campbell, D. T., & Stanley, J. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Connor, W. R. The effect of testing on learning in physics. Unpublished master's thesis, Univer. of Iowa, 1932. Cited by Kirkpatrick (1933).

Curo, D. M. An investigation of the influence of daily pre-class testing on achievement in high school american history classes. Unpublished doctoral dissertation, Purdue Univer., 1963.

Kitch, L. V. An experiment in integrating testing with learning in high school biology. Unpublished master's thesis, Univer. of Southern California, 1932.

McClymond, D. M. The relation of test items to textbook content in physics. Unpublished master's thesis, Univer. of Iowa, 1932. Cited by Kirkpatrick (1933).

Maloney, Estelle L., & Ruch, G. M. The use of objective tests in teaching as illustrated by grammar. Sch. Rev., 1929, 37, 62-66. Cited by Kirkpatrick (1933).

Pikunas, J., & Mazzota, D. The effects of weekly testing in the teaching of science. Sci. Educ., 1965, 49, 373-376.

Weissman, S. A. The effect of frequent testings on achievement in high school physics. Unpublished master's thesis, College of the City of New York, 1934. Cited by Gable (1936).

EXPERIMENTAL STUDIES: POST-HIGH SCHOOL

Deputy, E. C. Knowledge of success as a motivating influence in college work. J. Educ. Res., 1929, 20, 327-334.

Eurich, A. C., Longtaff, H. P., & Wilder, Marion. The effect of weekly examinations upon achievement in psychology. In The effective general college curriculum as revealed by examinations: a report of the committee on educational research of the university of minnesota. Minneapolis: Univer. of Minnesota, 1937. Pp. 333-347.

Fattu, N. A. Testing as a teaching device. High sch. J., 1957, 79-82.

Fitch, Mildred L., Drucker, A. J., & Norton, J. A., Jr. Frequent testing as a motivating factor in large lecture classes. J. educ. Psychol., 1951, 42, 1-20.

Güetzkow, H., Kelley, E. L., & McKeachie, W. J. An experimental comparison of recitation, discussion, and tutorial methods in college teaching. J. educ. Psychol., 1954, 45, 193-207.

Johnson, Bess E. The effect of written examinations on learning and on the retention of learning. J. exp. Educ., 1938, 7, 55-62. Cited by Mudgett (1956).

Keys, N. The influence on learning and retention of weekly as opposed to monthly tests. J. educ. Psychol., 1934, 25, 427-436. (a)

Kulp, D. H. II. Weekly tests for graduate students? Sch. & Soc., 1933, 38, 157-159.

Laidlaw, W. J. The effects of frequent tests on achievement, retention, and transfer, and test behavior. Unpublished doctoral dissertation, Columbia Univer., 1963.

Maize, R. C. Two methods of teaching english composition to retarded college freshmen. J. educ. Psychol., 1954, 45, 22-28. Cited by McKeachie (1963).

Mudgett, A. G. The effects of periodic testing on learning and retention in engineering drawing. Unpublished doctoral dissertation, Univer. of Minnesota, 1956.

Noll, V. H. The effect of written tests upon achievement in college classes: an experiment and a summary of evidence. J. educ. Res., 1939, 32, 345-358.

Ross, C. C., & Henry, L. K. The relation between frequency of testing and progress in learning psychology. J. educ. Psychol., 1939, 30, 604-611. Cited by Laidlaw (1963), Mudgett (1956), and Wrightstone (1963).

Selakovich, D. An experiment attempting to determine the effectiveness of frequent testing as an aid to learning in beginning college courses in american government. J. educ. Res., 1962, 55, 178-180.

Serenius, C. A. Objective drill material versus informal discussion in college history teaching. Unpublished master's thesis, Univer. of Iowa, 1930. Cited by Kirkpatrick (1933).

Smeltzer, C. H. An experimental evaluation of certain teaching procedures in educational psychology. Unpublished doctoral dissertation, Ohio State Univer., 1931.

Standlee, L. S., & Popham, W. J. Quizzes' contribution to learning. J. educ. Psychol., 1960, 51, 322-325.

Sumner, F. C., & Brooker, Nettie M. Prognostic and other values of daily tests. J. appl. Psychol., 1944, 28, 323-328.

Turney, A. H. The effect of frequent short objective tests upon the achievement of college students in educational psychology. Sch. & Soc., 1931, 33, 760-762.

REVIEW TOPIC TWO: TEST GRADES

Bostrum, R. N., Vlandis, J. W., & Rosenbaum, M. E. Grades as reinforcing contingencies and attitude change. J. educ. Psychol. 1961, 52, 112-115.

Fay, P. J. The effect of the knowledge of marks on the subsequent achievement of college students. J. educ. Psychol., 1937, 28, 548-554.

Hawk, T. L., & DeRidder, L. M. A comparison of the performance of pre-graded students with grade-motivated students. J. educ. Res., 1963, 548-550.

Kneeland, Natalie, & Bernard, Louise. Use objective "tests" to stimulate good discussion: especially in d. e. classes. Bus. educ.

World, 1953, 499-500.

Nolan, D. M. The experimental effect of grades assigned to a single task on subsequent academic performance. Unpublished doctoral dissertation, Michigan State Univer., 1964.

Odell, C. W. Traditional examinations and new-type tests. New York: Century, 1928. Pp. 17, 39, 56. Cited by Kirkpatrick (1933).

Panlasigui, I., & Knight, F. B. The effect of the awareness of success or failure. Yearb. nat. Soc. Stud. Educ., 1930, 29, Part II.

REVIEW TOPIC THREE: TEST CORRECTION

Buckner, C. A. Unpublished manuscript, 1931. Cited by Kitch (1932).

Cocks, A. W. The pedagogical value of the true-false examination. Univer. res. Monogr., 1929, No. 7.

Curtis, F. D. Testing as a means of improving instruction. Sci. Educ., 1944, 28, 29-31.

Curtis, F. D., & Darling, W. C. Teaching values of common practices in correcting examination papers--a second study. Sch Rev., 1932, 40, 515-525.

Curtis, F. D., & Woods, G. G. A study of the relative teaching values of four common practices in correcting examination papers. Sch. Rev., 1929, 37, 615-623.

Davis, R. A. Testing and the course of classroom learning. J. educ. Psychol., 1943, 34, 526-534.

Gates, A. I. The true-false test as a measure of achievement in college courses. J. educ. Psychol., 1921, 12, 276-287. Cited by

Cocks (1929).

Jeep, H. A. Must objective tests be dogmatic? Educ. Admin. Supervis., 1933, 19, 181-190. Cited by Lee & Symonds (1934).

Krause, Ruthetta. Paper grading and checking: what about student attitudes and responsibilities? Bus. educ. World, 1966, 47, 36, 43-45.

Lee, J. M., & Symonds, P. M. New type or objective tests: a summary of recent investigations. J. educ. Psychol., 1934, 25, 161-184.

Smeltzer, C. H. Educational engineering in testing and diagnosis. Educ. Meth., 1933, 12, 526-530. (a)

REVIEW TOPIC FOUR: TEST RESULT FEEDBACK

Anderson, S. B. Can tests teach? Cal. J. second. Educ., 1960, 35, 50-55.

Angell, G. W. Effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. J. educ. Res., 1949, 42, 391-394. Cited by Findley & Smith (1950).

Book, W. F., & Norvell, L. The will to learn: an experimental study in incentives to learning. Pedag. Sem., 1922, 29, 305-362. Cited by Deputy (1929) and Fitch et al. (1951).

Brown, F. J. Knowledge of results as an incentive in school room practice. J. educ. Psychol., 1932, 23, 532-552.

Bryan, G. L., & Rigney, J. W. An evaluation of a method for shipboard training in operations knowledge. Tech. Rep. 18, Electronics Personnel Res. Grp., Univer. of Southern California,

1956. Cited by McKeachie (1963).

Coladarci, A. P. Cited by Storey (1964).

Daniel, Kathryn B. The effects of immediate and delayed reinforcement with respect to knowledge of test results. Paper read at AERA convent., Chicago, February, 1968.

Daniel, Kathryn B., & Witchel, B. A comparison of immediate and delayed reinforcement groups with respect to retention of knowledge. Paper read at AERA convent., New York, February, 1967. Cited by Daniel (1968).

Dyer, H. S. Needed changes to sweeten the impact of testing. Personnel Guid. J., 1967, 776-780.

Flook, A. J. M. Note on the use of new-type tests for improving the quality of discussion in discussion groups. Brit. J. educ. Psychol., 1959, 29, 261-263.

Jones, H. L., & Sawyer, M. O. A new evaluation instrument. J. educ. Res., 1949, 42, 381-385. Cited by Findley & Smith (1950).

Koester, G. A. Using instructor-made tests for instructional purposes. Educ. res. Bull., 1957, 36, 207-208.

Krueger, W. C. Students' honesty in correcting grading errors. J. appl. Psychol., 1947, 31, 533-535. Cited by Findley & Smith (1950) and Mudgett (1956).

Opdyke, J. B. Constructive examinations. Educ. Rev., 1927, 73, 33-43. Cited by Kirkpatrick (1933).

Page, E. B. The effects upon student achievement of written comments accompanying letter grades. Unpublished doctoral dissertation, Univer. of California, Los Angeles, 1958. (a). Cited by Page (1958b).

Page, E. B. Teacher comments and student performance: a seventy-four classroom experiment in school motivation. J. educ. Psychol., 1958, 49, 173-181. (b)

Paige, D. D. Learning while testing. J. educ. Res., 1966, 59, 276-277.

Plowman, L., & Stroud, J. B. Effects of informing pupils of the consequences of their responses to objective test questions. J. educ. Res., 1942, 36, 16-20. Cited by Sassenrath & Garverick (1965).

Ross, C. C. The influence upon achievement of a knowledge of progress. J. educ. Psychol., 1933, 24, 609-619.

Sassenrath, J. M., & Garverick, C. M. Effects of differential feedback from examinations on retention and transfer. J. educ. Psychol., 1965, 56, 259-263.

Sturges, P. T., & Crawford, J. J. The effect of delay of reinforcement on learning factual material. Paper read at Western Psychol. Ass., Santa Monica, California, April, 1963. Cited by Daniel (1968).

Symonds, P. M. Measurement in secondary education. New York: Macmillan, 1927. Pp. 1,533. Cited by Kirkpatrick (1933).

REVIEW TOPIC FIVE: PRETESTING

Breed, F. S. How to teach spelling. Danville, N. J.: F. A. Owen, 1930. Cited by Kirkpatrick (1933).

Breslick, E. R. Testing as a means of improving the teaching of high school mathematics. Math. Teacher, 1921, 14, 277. Cited

by Landis (1928).

Burr, S. E. The value of examinations. In F. Cody et al. (1929). P. 227.

Campbell, D. T. Factors relative to the validity of experiments in social settings. Psychol. Bull., 1357, 54, 297-312.

Campbell, D. T. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Cody, F. The value of examinations. In F. Cody et al. (1929). P. 62.

Davis, R. A. Testing and the course of classroom learning. J. educ. Psychol., 1943, 34, 526-534.

Edling, J. V. A study of the effectiveness of audiovisual teaching materials when prepared according to the principle of motivational research. Final Report, Title VII, Project No. 221, Grant No. 735055.00.11. Washington: U. S. O. E., 1963. Cited by Rayder & Neidt (1964).

Entwisle, Doris R. Interactive effects of pretesting. Educ. psychol. Measmt., 1961, 21, 607-620.

Gates, A. I. An experimental comparison of the study-test and the test-study method in spelling. J. educ. Psychol., 1939 [?], 1-20.

Horn, E. Principles of method in teaching spelling as derived from scientific investigation. In Yearb. nat. Soc. Stud. Educ., 1919, 18, Part II. Pp. 52-73. Cited by Kirkpatrick (1933).

Horn, E., & Ashbaugh, E. J. Lippincott's new horn-ashbaugh speller. Philadelphia: Lippincott, 1926. Cited by Kirkpatrick (1933).

Hovland, C. I., Janis, I. L., & Kelley, H. H. Communication and persuasion. New Haven: Yale, 1953. Cited by Entwisle (1961).

Hutchinson, M. E. The function of examinations. Educ., 1933, 436-439.

Jersild, A. T. Examination as an aid to learning. J. educ. Psychol., 1929, 20, 602-609.

Jones, R. G. The value of examinations. In F. Cody et al. (1929). P. 119.

Keys, N. The influence of true-false items on specific learning. J. educ. Psychol., 1934, 25, 511-520. (b)

Kilzer, L. R. The study test method versus the test study method in teaching spelling. Unpublished master's thesis, Univer. of Iowa, 1926. Cited by Kirkpatrick (1933).

Kingsley, J. H. The test study method versus the study test method in spelling. Element. sch. J., 1923, 24, 126-129. Cited by Kirkpatrick (1933) and Luce (1939).

Kirkpatrick, J. E. Motivating effect of a specific type of testing program. Unpublished doctoral dissertation, State Univer. of Iowa, 1933.

Kneeland, Natalie, & Bernard, Louise. Use objective "tests" to stimulate good discussion : especially in d. e. classes. Bus. educ. World, 1953, 499-500.

Lana, R. E. A further investigation of the pretest-treatment interaction effect. J. appl. Psychol., 1959, 43, 421-422. (a). Cited by Entwisle (1961).

Lana, R. E. Pretest-treatment interaction effects in attitudinal

studies. Psychol. Bull., 1959, 56, 293-300. (b).

Lana, R. E., & King, D. J. Learning factors as determiners of pretest sensitization. J. appl. Psychol., 1960, 44, 189-191. Cited by Rayder & Neidt (1964).

Lazarsfeld. In Campbell (1957). Cited by Entwisle (1961).

Lockhart, Aileene. Testing can improve teaching. J. Health phys. Educ., 1948, 19, 590, 627-629.

Luce, Edna R. The effect of pre-tests and post-tests upon retention. Unpublished master's thesis, State Univer. of Iowa, 1939.

McGinnis, W. C. The value of examinations. In F. Cody et al. (1929). P. 63.

Piers, E. V. In Bull. maritime psychol. Ass., 1955, 53-56. (Psychol. Abstr., 30). Cited by Entwisle (1961).

Rayder, N. F., & Neidt, C. O. Attitude change as a function of the number of scales administered. AV commun. Rev., 1964, 12, 402-412.

Smeltzer, C. H. Educational engineering in testing and diagnosis. Educ. Meth., 1933, 12, 526-530. (a)

Solomon, R. L. An extension of control group design. Psychol. Bull., 1949, 46, 137-150.

Spencer, D. L. The improvement of teaching by means of "home-made" non-standard diagnostic tests and remedial instruction. Sch. Rev., 1923, 31, 276-281. Cited by Kirkpatrick (1933).

Tiedeman, H. R. A study in retention of classroom learning. J. educ. Res., 1948, 41, 516-531.

Watts, W. An investigation of the test study method of teaching spelling. Unpublished master's thesis, Univer. of Iowa, 1928. Cited by Kirkpatrick (1933) and Luce (1939).

Weber, H. C. The value of examinations. In F. Cody et al. (1929). Pp. 62-63.

REVIEW TOPIC SIX: RETESTING

Celinski, O. Announced repetitive tests as a basis for self-directed study and evaluation. J. exp. Educ., 1968, 36, 17-26.

Davis, R. A., & Rood, E. J. Remembering and forgetting arithmetical abilities. J. educ. Psychol., 1947, 38, 216-222. Cited by Findley & Smith (1950).

Little, E. B. Pre-test and re-test scores in retention calculation. J. exp. Educ., 1960, 29, 161-167.

Sones, A. M., & Stroud, J. B. Review, with special reference to temporal position. J. educ. Psychol., 1940 [?], 665-676.

Spencer, E. M. The retention of orally presented materials. Unpublished doctoral dissertation, State Univer. of Iowa, 1940.

Spitzer, H. F. A study of retention in reading. Unpublished doctoral dissertation, State Univer. of Iowa, 1938. Cited by Luce (1939) and Spencer (1940).

Spitzer, H. F. Studies in retention. J. educ. Psychol., 1939, 30, 641-656. Cited by Sones & Stroud (1940), Spencer (1940), and Tiedeman (1948).

Woodworth, R. S. Experimental psychology. New York: Holt, 1938. Cited by Spencer (1940).

Yoakam, G. A. The effect of a single reading on the retention of various types of materials in the content subjects of the elementary school curriculum as measured by immediate and delayed recall. Unpublished doctoral dissertation, State Univer. of Iowa, 1922.

REVIEW TOPIC SEVEN: TEST EXPECTATION

Corey, S. M. The effect of motivation upon the relationship between achievement and intelligence. Sch. & Soc., 1935, 41, 256-257. Cited by Gable (1936).

Gable, Sister Felicita. The effect of two contrasting forms of testing upon learning. The Johns Hopkins Univer. Studies in Educ., 1936, No. 25.

Jones, H. E. Experimental studies of college teaching: the effect of examination on permanence of learning. Arch. Psychol., N. Y., 1923, No. 68.

Pease, G. R. The effect of cramming upon retention, immediate and delayed. Unpublished master's thesis, State Univer. of Iowa, 1927. Cited by Pease (1930).

Pease, G. R. Should teachers give warning of tests and examinations? J. educ. Psychol., 1930, 21, 273-277.

Schutte, T. H. Is there value in the final examination? J. educ. Res., 1925, 12, 204-213.

White, H. B. Testing as an aid to learning. Educ. Admin. Supervis., 1932, 18, 41-46.

REVIEW TOPIC EIGHT: TEST EXEMPTION

Cole, T. R. The value of examinations. In F. Cody et al. (1929). P. 120.

Davis, R. A. Testing and the course of classroom learning. J. educ. Psychol., 1943, 34, 526-534.

Dole, A. A. Evidence of the effectiveness of a program for giving college credits by examination. Educ. psychol. Measmt., 1951, 11, 387-395. Cited by Gardner (1953).

Engelhart, M. D. The effect of exemption from final examination on the distribution of term grades. J. educ. Res., 1931, 23, 319-321.

Gould, G. Practices in marking and examination. Sch. Rev., 1932, 40, 142-146.

Meltzer. Unpublished manuscript. In S. L. Pressey, Psychology and the new new education. New York: Harper, 1933. Pp. 363-366. Cited by Laidlaw (1963) and Noll (1939).

Morley, E. E. Final examinations and the effect of exemptions. High sch. Teacher, 1926, 2, 90-91. Cited by Engelhart (1931).

Nickerson, Fern. The value of examinations. In F. Cody et al. (1929). P. 253.

Odell, C. W. Traditional examinations and new-type tests. New York: Century, 1928. Pp. 17, 39, 56. Cited by Kirkpatrick (1933).

Remmers, H. H. Exemption from college semester examinations as a condition of learning. Purdue Univer. Bull., Studies in Higher Educ., 1933, 134, No. 3. Cited by Mudgett (1956).

Smeltzer, C. H. An experimental evaluation of certain teaching procedures in educational psychology. Unpublished doctoral dissertation, Ohio State Univer., 1931.

Smeltzer, C. H. Improving and evaluating the efficiency of college instruction. J. educ. Psychol., 1933, 24, 282-302. (b) Cited by Mudgett (1956).

Webb, J. C. The value of examinations. In F. Cody et al. (1929). Pp. 281-282.

REVIEW TOPIC NINE: STUDENT PREPARATION FOR TESTS

Class, E. C. The effect of the kind of test announcement on student's preparation. J. educ. Res., 1935, 28, 358-361. Cited by Mudgett (1956).

Cody, F., Weber, H. C., McGinnis, W. C., Pearson, M. E. Jones, R. G., Pyrtle, E. Ruth, Holmes, S. H., Wofford, Kate V., Chewning, J. O., Cole, T. R., Lund, J., Geiger, W. F., Ballou, F. W., Slawson, S. J., Burr, S. E., Threlkeld, A. L., Adams, W. C. T., Nickerson, Fern, Weeks, I. D., Webb, J. C., Venables, M. C., & Weber, S. E. The value of examinations : a symposium. J. Educ., 1929, 62-63, 119-120, 226-227, 252-253, 281-282, 536-537.

Cole, T. R. The value of examinations. In F. Cody et al. (1929). P. 120.

Colvin, S. S. The learning process. New York:Macmillan, 1913. P. 176.

Douglass, H. R., & Tallmadge, Margaret. How university students prepare for new types of examinations. Sch. & Soc., 1934, 39, 318-320. Cited by Tyler (1959) and Wrightstone (1963).

Krause, Ruthetta. Paper grading and checking: what about student attitudes and responsibilities? Bus. educ. World, 1966, 47, 36, 43-45.

Meyer, G. Experimental study of old and new types of examination. J. educ. Psychol., 1934, 25, 641-661, and 1935, 26, 30-40. Cited by Tyler (1959) and Wrightstone (1963).

Pearson, M. E. The value of examinations. In F. Cody et al. (1929). P. 63.

Pyrtle, E. Ruth. The value of examinations. In F. Cody et al. (1929). P. 119.

Ruch, G. M. The objective or new-type examination. Chicago: Scott, Foresman, 1929. Pp. 10, 25, 145. Cited by Kirkpatrick (1933) and Turney (1931).

Symonds, P. M. Measurement in secondary education. New York: Macmillan, 1927. Pp. 1, 533. Cited by Kirkpatrick (1933).

Tyler, R. W. What testing does to teachers and students. In Invitat. Conf. Test. Prob. Proc. Princeton: Educational Testing Service, 1959. Pp. 10-16.

Vallance, T. R. A comparison of essay and objective examinations as learning experiences. J. educ. Res., 1947, 41, 279-288. Cited by Findley & Smith (1950).

Venables, M. C. The value of examinations. In F. Cody et al. (1929). P. 282.

Weber, H. C. The value of examinations. In F. Cody et al. (1929). Pp. 62-63.

Weeks, I. D. The value of examinations. In F. Cody et al. (1929). P. 281.

REVIEW TOPIC TEN: STUDENT ATTITUDES TOWARD TESTS

Bender, W., Jr., & Davis, R. A. What high school students think about teacher-made examinations. J. educ. Res., 1949, 43, 58-65.

DeLong, A. R. Emotional effects of elementary school testing. Understanding the Child, 1955 [?], 103-107.

Deputy, E. C. Knowledge of success as a motivating influence in college work. J. educ. Res., 1929, 20, 327-334.

Gaier, E. I. Students' perceptions of factors affecting test performance. J. educ. Res., 1962, 55, 561-566.

Koester, G. A. Using instructor-made tests for instructional purposes. Educ. res. Bull., 1957, 36, 207-208.

Turney, A. H. The effect of frequent short objective tests upon the achievement of college students in educational psychology. Sch. & Soc., 1931, 33, 760-762.

REVIEW TOPIC ELEVEN: TEST TYPE

Ballard, F. B. The new examiner. London: Hodder & Stoughton, 1925. Cited by Cocks (1929) and Keys (1934b).

Cocks, A. W. University research monographs: no. 7: the pedagogical value of the true-false examination. Baltimore: Warwick & York, 1929.

Jersild, A. T. Examination as an aid to learning. J. educ. Psychol., 1929, 20, 602-609.

McCall, W. A. A new kind of school examination. J. educ. Res., 1920, 1, 33-46. Cited by Cocks (1929).

Remmers, H. H., & Remmers, Edna M. The negative suggestion effect of true-false examination questions. J. educ. Psychol., 1925 [?], 52-56.

REVIEW TOPIC TWELVE: "TEST-LIKE EVENTS"

Anderson, F. C. An experiment to determine the effect of a work book on achievement in general science. Unpublished manuscript. (Rev. educ. Res., 1942, 12, 383). Cited by McKeachie & Hiler (1951).

Bruning, R. H. Effects of review and testlike events within the learning of prose materials. J. educ. Psychol., 1968, 59, 16-19.

Cason, H. An intelligent-question method of teaching and testing. J. genet. Psychol., 1939, 54, 359-390. Cited by McKeachie & Hiler (1951).

Greene, E. B. Certain aspects of lecture, reading and guided reading. Sch. & Soc., 1934, 39, 619-624. Cited by McKeachie & Hiler (1951).

Harrington, E. E., & Lippert, D. E. The work book. Penna. sch. J., 1934, 82, 359-362. Cited by McKeachie & Hiler (1951).

Hertzberg, O. E., Heilman, J. D., & Leuenberger, H. W. The value of

objective tests as teaching devices in educational psychology classes. J. educ. Psychol., 1932, 23, 371-380.

Hurd, A. W. The textbook versus work sheets in instruction. Educ. Admin. Supervis., 1931, 17, 661-664. (a). Cited by Kirkpatrick (1933).

Hurd, A. W. The workbook as an instructional aid. Sch. Rev., 1931, 39, 608-616. (b). Cited by Kirkpatrick (1933) and McKeachie & Hiler (1951).

Langman, Muriel P. Set, attention, and purpose in reading. Educ., 1963, 83, 532-536.

Motion Picture Research Project. Do "motivation" and "participation" questions increase learning? Educ. Screen, 1947, 256-259, 274, 283.

McKeachie, W. J., & Hiler, W. The problem-oriented approach to teaching psychology. J. educ. Psychol., 1951 [?], 224-232.

Robinson, L. J. The value to college students of lists of questions on a text. Unpublished master's thesis, Colorado State Teachers College, 1926. Cited by Hertzberg et al. (1932).

Rothkopf, E. Z. Some conjectures about inspection behavior in learning from written sentences and the response mode problem in programed self-instruction. J. programed Instruction, 1963, 2, 31-45.

Rothkopf, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand-McNally, 1965. Pp. 193-221.

Rothkopf, E. Z. Calculus of practice and mathemagenic behavior: two research approaches to the scientific management of the instructional process. Paper read at Phi Delta Kappan Research Symposium, Berkeley, Cal., November 28-29, 1966. (a)

Rothkopf, E. Z. Concerning parallels between adaptive processes in thinking and self-instruction. Paper read at Univer. of Pittsburgh Symposium on Approaches to Thought, Pittsburgh, October 13-14, 1966. (b)

Rothkopf, E. Z. Learning from written instructive materials: an exploration of the control of inspection behavior by test-like events. Amer. educ. res. J., 1966, 3, 241-249. (c)

Rothkopf, E. Z. Textual constraint as function of repeated inspection. J. educ. Psychol., 1968, 59, 20-25.

Rothkopf, E. Z., & Bisbicos, Ethel E. Selective facilitative effects of interspersed questions on learning from written materials. J. educ. Psychol., 1967, 58, 56-61.